Multilocus Phylogeny Estimation Using Probabilistic 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, 26 Oxford Street, Cambridge, MA 02138, USA*

**Corresponding author: Marzieh Khodaei, email: mk16e@fsu.edu*

Abstract

Inferring the evolutionary history of species or populations employing multilocus analysis is gaining ground in phylogenetic analysis. We developed an alignment-free method to infer the multilocus species tree, which is implemented in the Python package TopicContml. The method operates in two primary stages. First, it uses probabilistic topic modeling (specifically, Latent Dirichlet Allocation or LDA) to extract topic frequencies from k-mers, which are in turn derived from multilocus DNA sequences. Second, these extracted frequencies serve as an input for the program Contml in the PHYLIP package, which is used to generate a species tree. We evaluated the performance of our method with biological and simulated datasets: a dataset with 14 DNA sequence loci from 78-92 individuals from two Australian bird species distributed in 9 populations, and a second dataset of 67317 autosomal loci and 4157 X-chromosome loci of 6 species in the *Anopheles gambiae* complex, and several simulated data sets. Our empirical results and simulated data suggest that our method is efficient and statistically accurate. We also assessed the uncertainty of the estimated relationships among clades using a bootstrap procedure for aligned data and for k-mers.

Key words: topic modeling; NLP; LDA; ContML; multilocus analysis; k-mers; alignment-free; bootstrap.

Introduction

Recently, alignment-free approaches based on shared properties of sub-sequences of defined length k (k-mers or k-grams) (Marçais and Kingsford 2011) have been developed to compare sequences. Alignment-free methods for evolutionary analysis have been reviewed (Vinga and Almeida 2003; Zielezinski et al. 2019) and the robustness of them were investigated (Chan et al. 2014; Bernard et al. 2016). For example, they can be used to derive distances to be summarized into phylogenies (Edwards et al. 2002; Chan et al. 2014). Alignment-free approaches for phylogeny inference have attracted considerable attention in recent years (Aledo 2022; Dylus et al. 2023). A number of studies have shown that the alignment artifacts can have a significant impact on tree topology (Ogden and Rosenberg 2006; Wong et al. 2008). Therefore, alignment-free approaches could provide a promising alternative to overcome the weaknesses of alignment-based methods (Ren et al. 2018).

Probabilistic topic modeling (Blei et al. 2003), is a statistical approach whose goal is to identify major themes and connections among themes in documents and other large collections of text. It originated from the field of Natural Language Processing (NLP) and was introduced by statisticians looking for applications of machine learning. Griffiths and Steyvers (2004) applied this method to infer topics from a large number of scientific documents, called a corpus. Latent Dirichlet Allocation (LDA) is a popular technique in topic modeling introduced by Blei et al. (2003) within the context of unsupervised machine learning. Recently, LDA has also been the focus of much attention within the bioinformatics community, and various applications to biological data have been researched and analyzed (Liu et al. 2016). Recently, some applications of alignment-free methods have been presented to solve problems offered by DNA sequences or genomes. For example, in a statistical application, LDA was used by La Rosa et al. (2015) to extract the frequency of

fixed-length k-mers (words) of DNA sequences (documents) and thereby discover latent patterns in massive biological data to be used for clustering and classification of DNA sequences. Other studies have adopted LDA clustering for analysis of single cells expression or epigenetic data (duVerle et al. 2016; Dey et al. 2017). Here, we adopt probabilistic topic modeling to present a new computational approach to construct evolutionary relationships among individuals of different populations or species using multilocus data that may contain unaligned or aligned DNA sequences. We will use the term species tree for these trees, whether they are derived from individuals grouped into populations or species.

Multilocus datasets (sequences of multiple genes per species) are becoming increasingly prominent in inferring the evolutionary history of closely related species. Two analytical methodologies to analyse multilocus datasets are (1) concatenation methods in which all gene sequences of each species are concatenated and the generated single alignment is used to generate a phylogenetic tree with programs such as RAxML (Stamatakis 2014) and IQ-TREE (Nguyen et al. 2015; Minh et al. 2020); (2) coalescent-based approaches, for example, STAR (Liu et al. 2009b), SVDquartets (Chifman and Kubatko 2014), and ASTRAL-III (Zhang et al. 2018), in which genes are allowed to have different gene trees and then all those gene trees are combined into a species tree (Liu et al. 2009a; Mirarab et al. 2016). Our approach falls into group (2). We use probabilistic topic modeling to construct a species tree from multi-locus data consisting of unaligned or aligned DNA sequences. Our method is implemented in the Python package TopicContml. There are two key workflow phases: we first extract the topic frequencies of sequences using LDA, and then we use the program Contml (part of PHYLIP; Felsenstein 1981; 2004), Contml (Gene Frequencies and Continuous Characters Maximum Likelihood method), to estimates a phylogeny from frequency data using the restricted maximum likelihood method (REML) (Felsenstein 1973) based on the Brownian motion model for allele frequencies (Cavalli-Sforza and Edwards 1967).

The most widely used statistical method for assessing confidence in phylogenetic

analyses is the nonparametric bootstrap method (Efron 1979) introduced to the phylogenetics by Felsenstein (1985). Given a multiple sequence alignment, the bootstrap method involves resampling the original dataset with replacement of the aligned sites and creating the phylogeny for each replicate. For unaligned data, we use the approach used in Edwards et al. (2002), where a random sample of x k-mers is drawn from all x k-mers collected from the data. The fraction of the time a particular clade appears in the resulting bootstrap trees presents support values for the clades in the reference tree or in a majority-rule consensus tree (Holder et al. 2008; Efron et al. 1996; Hillis and Bull 1993). TopicContml implemented bootstrapping strategies for both aligned and unaligned sequences. We compare the TopicContml phylogenetic bootstrap with the resulting one of SVDquartets (Chifman and Kubatko 2014) implemented in Paup* (Swofford 2003) (SVDquartets +Paup*) which has been shown to be computationally efficient and be able to perform well under a range of scenarios (Chou et al. 2015).

Description

TopicContml *Software*

TopicContml is a Python package that is based on a two-phase pipeline: (1) it learns a probabilistic topic model from a multi-locus dataset of DNA sequences and extracts the topic frequencies of sequences using the LDA model for each locus (Fig. 1 left). For the analysis of data with multiple individuals from the same species or population, we set the option `--merging` n with the purpose of merging individual labels that start with the same n letters into groups. (2) It uses these topic frequencies from multiple loci to estimate a species tree with Contml (Fig. 1 right).

K-mer Decomposition.— K-mers are k-length substrings of a DNA or amino acid sequence. We decompose the DNA sequence into non-overlapping k-mers, as shown in Fig. 1. The current program uses a set of four different k of 2, 4, 6, and 8, which result in

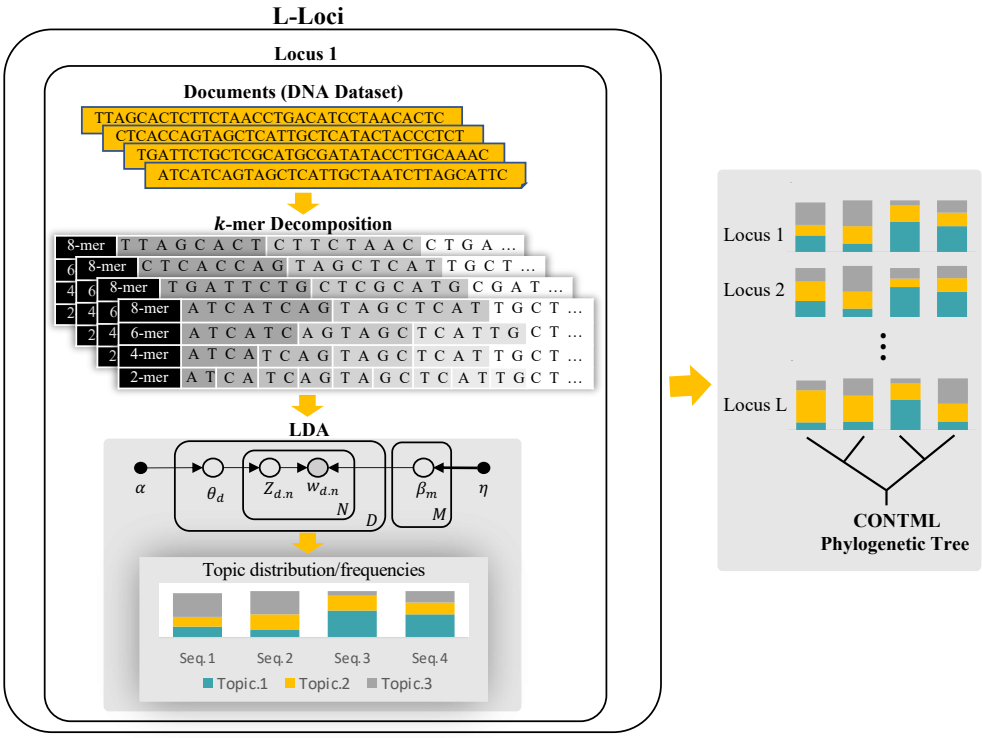

FIGURE 1. TOPICCONTML workflow to generate topic frequencies and the corresponding phylogeny.

more informative topic frequencies than overlapping or single k-mers.

Topic Modeling.— Given a collection of D documents and a number of M topics, topic modeling discover the M topics from a collection of text data, and estimates the probability of each topic for each document. We use LDA to extract these frequencies.

LDA *(Latent Dirichlet Allocation).—* The DNA sequences are assumed to be the corpus with D documents, and the A, T, C, G nucleotides are considered as letters. All DNA sequences are then decomposed into k-mers (words), resulting in a total number of V unique words. For each document $d \in D$, N is the number of words in the document, and $z_{d,n}$ indicates the topic for $w_{d,n}$ (the nth word in the dth document). θ_d is the distribution of topics for document d such that $\theta_d \sim Dir(\alpha)$, which is the Dirichlet distribution with parameter vector $\alpha > 0$, and β_m is the distribution of words for topic m such that

$\beta_m \sim Dir(\eta)$ with parameter vector $\eta > 0$. The only observed variables are the words w in the documents. The topics z for all words, the distribution θ over topics for all documents and the distribution β of words per topic are all latent variables in this model. The joint distribution is defined as

$$p(w, z, \theta, \beta|\alpha, \eta) = \prod_m p(\beta_m|\eta) \prod_d \left[p(\theta_d|\alpha) \prod_n p(z_{d,n}|\theta_d) p(w_{d,n}|z_{d,n}, \beta) \right]$$

Using the joint distribution, one can compute the posterior distribution of unknown model parameters, $p(\beta, \theta, z|w)$, which is the central task of topic modeling. We estimate the topic frequencies for every document, θ, using the Python package GENSIM (Řehůřek and Sojka 2010), these frequencies will then be used for the tree generation.

CONTML *(Gene Frequencies and Continuous Characters Maximum Likelihood method).*— The primary assumption of CONTML is that each character (each topic in our case) evolves independently according to a Brownian motion process and character changes since divergence are independent of each other, which means the net change after t units of time is normally distributed with zero mean and variance ct (the same constant c for all characters).

BENCHMARKS

Data Sets

We evaluate the accuracy of TOPICCONTML in two settings: real biological data and simulated data. We used two biological dataset that have very different features. A 14 locus data set of nine populations to discuss the accuracy of estimation, the effect of alignment, bootstrapping, and comparison to SVDQUARTETS, and a 7-species dataset of 67316 autosomal loci and 4173 X-chromosome loci to discuss the speed of operation. The simulated data was used to show the effects of the number of loci and the divergence times within the true genealogy of 40 individuals from 4 species or populations.

Real data set.— The data of nine populations of two bird species, Australian brown and black-tailed treecreepers (Climacteris picumnus and melanurus, respectively), were previously published by Edwards et al. (2023). The DNA sequences consist of 14 loci and 9 different geographic locations. For each locus, sequence length varied from 288 to 418 base pairs, and the number of sequences per species varied from 78 to 92 individuals. The sequences are aligned. In order to get unaligned sequences, we removed all gaps in each sequence. For each locus, we applied LDA for each locus across all 9 locations and extracted the topics frequencies. These topic frequencies of the 14 loci were used in ContML to generate the tree.

The second biological dataset was described in (Fontaine et al. 2015; Wen et al. 2016). We used the genome data of 6 species in the *Anopheles gambiae* complex and an outgroup from the DataDryad repository (Wen, D.; Yu, Y.; Hahn, M. W.; Nakhleh, L. [2016]. Data from: Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. Dryad. https://doi.org/10.5061/dryad.tn47c). the dataset contained 67,317 autosomal loci and 4173 X-chromsome loci, each locus with different sequence lengths, for autosomes: average 2299 base pairs with a range of 101 to 44,205 and for X: average 1167 base pairs with a range of 100 to 8010.

Simulated data set.— We simulated sequence data for 10 individuals from each of 4 populations, which were derived from ancestral populations according to the Newick tree structure: ((A,C),(B,D)). The time of split between (A , C) and (B , D) is $\tau/2$ and the root connecting the clades A,C and B,D was set at time τ (Fig. 2). The divergence time (in coalescent units) was varied: $\tau = [0.0, 0.01, 0.05, 0.1, 0.2]$, respectively. In each setting, 100 loci were simulated. Each sequence had a length of 1000bp. We used the coalescent simulator *ms* (Hudson 2002) to simulate the relationship among the individuals and the program *migdata* (https://github.com/pbeerli/popsimulate) to generate the DNA sequences.

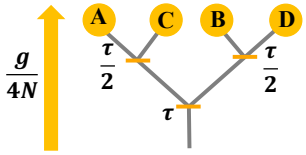

FIGURE 2. Population tree ((A,C),(B,D)) with split times. This tree was used to simulate trees with 10 individuals per population. The time scale is coalescent units, g is generations and N is the population size.

Multilocus Bootstrapping

To assess the statistical confidence of the inferred phylogenies, we use bootstrapping. TOPICCONTML has two options to generate bootstrap replicates: (1) unaligned sequences: in each locus, we chose x k-mers randomly in each locus and each individual (document) with x k-mers, and then generated bootstrap-replicate topic frequencies with the newly generated corpus of words; (2) aligned sequences: we chose randomly x site-column in the alignment to regenerate the new locus, and then used this bootstrap replicate to create k-mers and then topic frequencies per locus.

We constructed a majority-rule consensus tree from 1000 bootstrap trees using SUMTREES in DENDROPY (Sukumaran and Holder 2010). We also mapped the bootstrap replicates onto the best topology tree found by TOPICCONTML, as the reference tree, using SUMTREES.

We compared our species trees with SVDQUARTETS +PAUP*. This software infers species trees under the multispecies coalescent model. We assessed support with 1000 bootstrap replicates. Since the input format of SVDQUARTETS involves merging loci sequences, we first generated a merged NEXUS file of all individuals for all loci. For the missing sequences in some of the loci, we filled them with "?".

Results

Analysis of Real Data: Two species populations with multiple populations

For the bird dataset, we tokenized each DNA sequence at every locus using k-mer representation, with k values of 2, 4, 6, and 8, and non-overlapping tokens. We then merged the words from the same location and applied LDA on the corpus to generate the matrix of topic frequencies for the corresponding locus. The number of topics is an arbitrary constant in LDA. We explored a range of topics from 2 to 20; picking five topics led to results that recover the geographic relationship within a species well; by contrast, lower numbers of topics do not yield good resolution, and higher numbers do not improve the results. All analyses used five topics per locus. The computation time was only about one second per locus. Subsequently, we used the resulting matrix of multilocus topic frequencies to construct a maximum-likelihood tree with ContML in the PHYLIP package. The evaluation time of ContML depends on the number of tips and the number of loci. Evaluating 14 loci for 9 locations took only a few seconds whereas analyzing all 78 individuals as distinct 'species' took several minutes (results not shown). Fig. 3(b) and Fig. 3(c) show the trees generated by TopicContML for the 9 locations at which the birds were collected for unaligned and aligned sequences, respectively. Inspection of the clades reveals that our trees recover the geographic relationship within each species, and the locations are separated by species. Aligned and non-aligned trees are similar although some locations, for example, the locations Sedan, Bourke, and Noonbah are placed differently in both trees.

Bootstrapping.— To compare the performance of TopicContML with SVDquartets, we applied both methods to the aligned real dataset and reported 50% majority-rule consensus trees. For comparison, we also present the un-aligned bootstrap majority consensus tree. For each bootstrapped dataset, we followed the same pipeline described in the TopicContML Software section, generating topic frequencies and

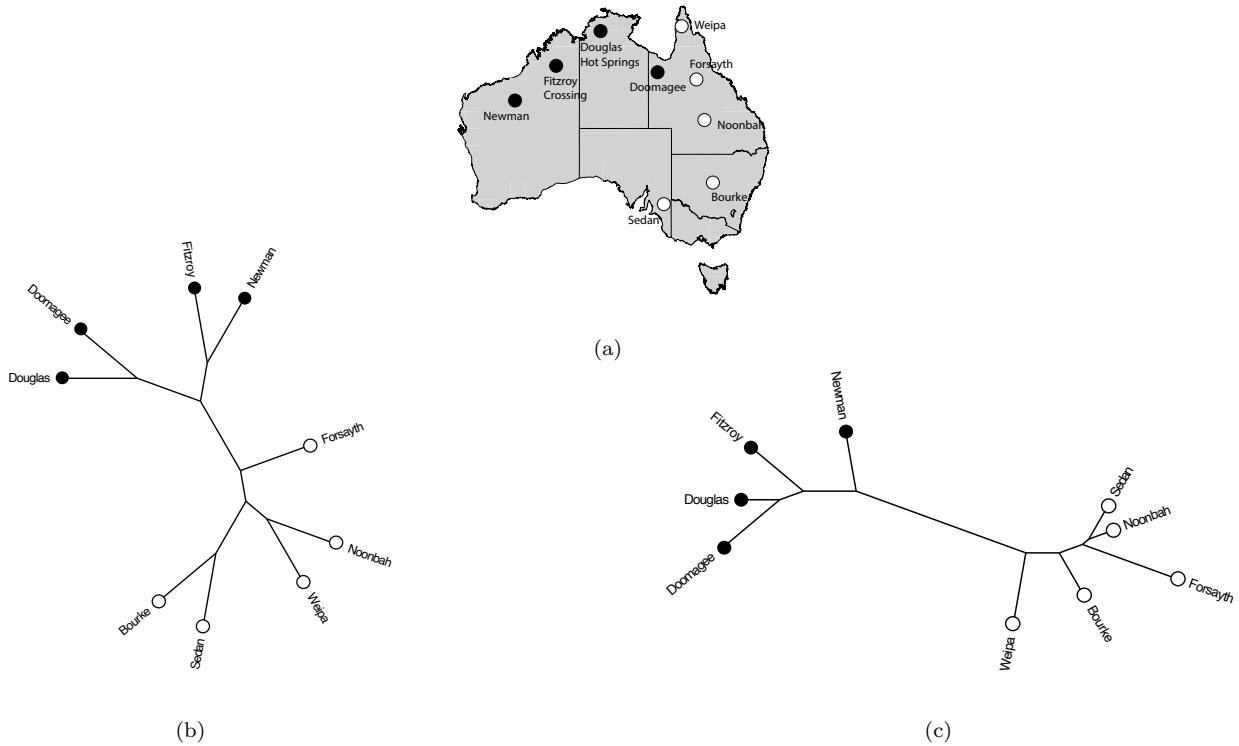

FIGURE 3. Relationship tree of 9 populations of two Australian treecreeper species reconstructed using TOPICCONTML. (a) The map of Australia shows the locations; the black disks mark *Climacteris melanurus*, and the white disks mark *Climacteris picumnus* (Edwards et al. 2023). (b) The phylogeny constructed by TOPICCONTML for unaligned sequences. (c) The phylogeny constructed by TOPICCONTML for aligned sequences.

corresponding phylogenies. Fig. 4 shows our results of bootstrapping with 1000 replicates. There are conspicuous differences between SVDQUARTETS and TOPICCONTML results. Comparing them to the mapping of the locations on Fig. 3(a), we observe that our bootstrap trees recover the relationship better or equal compared to the SVDQUARTETS method. Both methods have difficulty resolving some of the population splits, but are confidently separating the the two species.

Fig. 4 shows a comparison of TOPICCONTML bootstrap results with SVDQUARTETS +PAUP*. Both methods deliver similar bootstrap support values for the aligned data, TOPICCONTML support with unaligned data are not identical to the aligned data but recovers, in general, the geographic pattern of the locations well. The key difference between aligned and unaligned analyses seems to be that more potential splits with the unaligned data are unresolved.

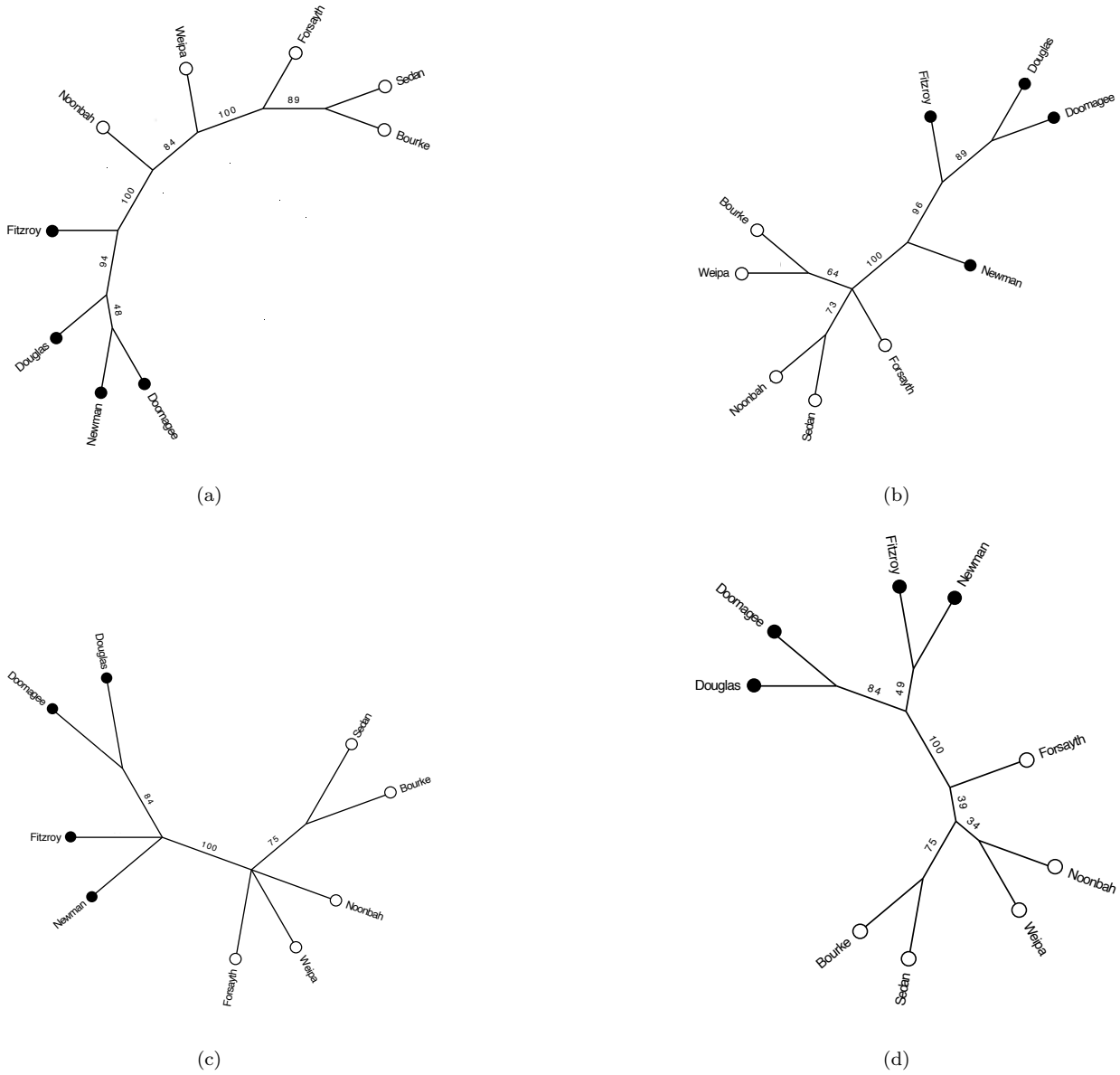

Figure 4. Bootstrap support comparison. (a) majority-consensus tree of aligned data analyzed by SVDquartets +Paup*. (b) majority-rule consensus tree of aligned data by TopicContml. (c) majority-rule consensus tree of unaligned data by TopicContml. (d) the best tree of unaligned data by TopicContml with bootstrap support superimposed. For each bootstrap analysis, 1000 replicates were used. Values in the graphs are % support.

Analysis of Real Data: Multilocus species tree from many loci

Recently, it became common to analyze multilocus datasets to derive species trees; several methods are available, but all will take a considerable amount of time to generate a tree and as Fontaine et al. (2015) and a reanalysis (Wen et al. 2016) of that data shows that a tree may not be the best way to visualize the relationships among the species. We evaluated 67517 autosomal loci of their *Anopheles gambiae* complex dataset; our analysis derived the same topology (Supplement Fig. S1(a)) as the whole-genome tree in Fig. 1B in Fontaine et al. (2015); this tree does not reflect the species phylogeny because of wide-spread introgression; in contrast, our analysis of the 4157 X-chromosome loci reflects the species tree reported in Fontaine et al. (2015) (Supplement Fig. S1(b)). Our runs of the aligned autosome and X-chromosome data took 22:34 minutes and 42 seconds, respectively. Both runs spend most of the time generating the topic frequencies (roughly about 1 to 2 seconds per 100 loci). In contrast, the unaligned data took roughly about 10× longer, because the unaligned data leads to more diverse k-mers. For example, the first 10 loci of the autosomes have on average 100 tokens for the data with gaps, and 514 tokens for the data with removed gaps. This increase in the number of tokens will then take longer to generate the topic frequencies and also the tree. The ContML analysis time of the seven tips was negligible. All our runs were done on an Apple Macbook Pro with an M1 Max (10 cores) using standard parallelization offered in the python module *multiprocessing*, suggesting that even very large datasets can be analyzed in a very short time on off-the-shelf computing devices.

Analysis of Simulated Data

The 4-population simulated data were analyzed similarly to the real data. We assessed the performance of TopicContML by counting the number of times the tree generated by TopicContML had the same topology as the true tree. Table 1 shows the results for different numbers of loci for five different sets of simulation data.

Table 1. Simulated 4-population trees: TopicContml tree congruence with the true topology: Percentage of agreement using Robinson-Foulds distance; Simulation parameters for the species trees: population size $\Theta = 100$ (4 $\times$ effective population size $\times$ mutation rate per locus); the split times are specified in coalescent units; and the number samples for each population was 10.

Loci	population split time* for population A,C and B,D				
	0.0	0.005	0.025	0.05	0.1
2	29	60	86	90	94
5	38	69	99	100	100
10	35	89	100	100	100
20	30	96	100	100	100
50	37	100	100	100	100
100	35	100	100	100	100

* Split time is in coalescent units (Fig. 2)

Our analyses show that the TopicContml method performs well in achieving topological similarity with the true tree topology. The more loci and the older the divergence time between the populations, the more often the true tree and the estimated tree have the same topology. With a population split time of zero, the simulation created a star phylogeny for the species. A 4-tip tree has three unrooted trees. Because there is no information about the split times, each of the three possible solutions should be around 33%: the average over all experiments in the table is 34%.

Discussion

This study demonstrates an integration of k-mers and probabilistic topic modeling for phylogenetic analysis of unaligned or aligned multilocus sequence data. The Python code TopicContml offers an efficient workflow to reconstruct evolutionary relationships without prior sequence alignment. TopicContml contains a two-phase analysis. First, k-mers are extracted from DNA sequences and Latent Dirichlet Allocation uses these k-mers to establish how probable an individual's set of k-mers fits several topics. For every locus and every individual, we generate a vector of assignment frequencies for a set of

topics. This step usually takes only seconds on modern computer architectures; it also can be parallelized among loci. The second step uses the topic frequencies as input for the program ContML to generate a phylogenetic tree. The evaluation time of ContML depends more on the number of tips than the number of loci. Our analyses delivered answers within seconds to minutes. The current version of ContML does not use any parallelization, but this could be implemented and would improve its runtime even for many species.

Testing on empirical genetic data and simulated data revealed that TopicContML recovers accurate tree topologies, particularly as more loci are included. The method demonstrated good performance even for difficult scenarios like recent divergence times. It also allows the assessment of tree uncertainty through bootstrap resampling. Compared to the well-known tool SVDquartets that is dependent on aligned data, bootstrap support was largely similar between the methods for aligned sequences. The trees derived from the unaligned datasets show longer terminal branches and shorter internal branches than when derived with aligned data; for the bootstrap, we have many trees with slightly modified sets of k-mers, leading to lower support for all clades than with aligned data, as expected, because the alignment enforces similarities among the individual sequences; therefore the k-mers that are derived from these aligned sequences are more similar. The very large number of autosomal loci for the mosquito dataset also revealed that unaligned data increases the number of unique tokens for each individual leading to longer run times and also leading to longer branch lengths. Additionally, this dataset also showed that TopicContML suffers the same issues as other tree-generating software in that introgression between different species leads to distorted trees. TopicContML is modular, and we have started to work on replacing ContML with a network-generating package that may improve the analyses of such datasets. For many datasets that do not suffer widespread introgression TopicContML proves a valuable contribution to the computational toolkit for phylogenetics by constructing evolutionary trees without or with

sequence alignment.

Software Availability

We implemented our method as a free software named TopicContml under the MIT open-source license. The source code and the documentation of TopicContml are available at https://github.com/TaraKhodaei/TopicContml.git

Acknowledgements

This research was partly supported by the National Science Foundation grant DBI2019989 to PB.

References

J. C. Aledo. Phylogenies from unaligned proteomes using sequence environments of amino acid residues. *Sci Rep*, 12(1):7497, 2022.

G. Bernard, C. X. Chan, and M. A. Ragan. Alignment-free microbial phylogenomics under scenarios of sequence divergence, genome rearrangement and lateral genetic transfer. *Sci Rep*, 6(1):28970, 2016.

D. M. Blei, A. Y. NG, and M. I. Jordan. Latent Dirichlet allocation. *J Mach Learn Res*, 3: 993–1022, 2003.

L. L. Cavalli-Sforza and A. W. F. Edwards. Phylogenetic analysis. Models and estimation procedures. *Am J Hum Gen*, 19(3 Pt 1):233–257, 1967.

C. X. Chan, G. Bernard, O. Poirion, J. M. Hogan, and M. A. Ragan. Inferring phylogenies of evolving sequences without multiple sequence alignment. *Sci Rep*, 4(1):6504, 2014.

J. Chifman and L. Kubatko. Quartet inference from SNP data under the coalescent model. *Bioinformatics*, 30(23):3317–3324, 2014.

J. Chou, A. Gupta, S. Yaduvanshi, R. Davidson, M. Nute, S. Mirarab, and T. Warnow. A comparative study of SVDquartets and other coalescent-based species tree estimation methods. *BMC genomics*, 16(10):1–11, 2015.

K. K. Dey, C. J. Hsiao, and M. Stephens. Visualizing the structure of RNA-seq expression data using grade of membership models. *PLoS Genet*, 13(3):e1006599, 2017.

D. A. duVerle, S. Yotsukura, S. Nomura, H. Aburatani, and K. Tsuda. CellTree: an R/bioconductor package to infer the hierarchical structure of cell populations from single-cell RNA-seq data. *BMC Bioinformatics*, 17(1):363, 2016.

D. Dylus, A. Altenhoff, S. Majidian, F. J. Sedlazeck, and C. Dessimoz. Inference of phylogenetic trees directly from raw sequencing reads using Read2Tree. *Nat Biotechnol*, pages 1–9, 2023.

S. V. Edwards, B. Fertil, A. Giron, and P. J. Deschavanne. A genomic schism in birds revealed by phylogenetic analysis of DNA strings. *Syst Biol*, 51(4):599–613, 2002.

S. V. Edwards, J. F. R. Tonini, N. Mcinerney, C. Welch, and P. Beerli. Multilocus phylogeography, population genetics and niche evolution of Australian brown and black-tailed treecreepers (Aves: Climacteris). *Biological Journal of the Linnean Society*, 138(3):249–273, 2023.

B. Efron. Bootstrap methods: another look at the jackknife. *Ann Statis*, 7(1):1–26, 1979.

B. Efron, E. Halloran, and S. Holmes. Bootstrap confidence levels for phylogenetic trees. *Proc Natl Acad Sci*, 93(23):13429–13429, 1996.

J. Felsenstein. Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am J Hum Gen*, 25(5):471–492, 1973.

J. Felsenstein. Evolutionary trees from gene frequencies and quantitative characters: finding maximum likelihood estimates. *Evolution*, pages 1229–1242, 1981.

J. Felsenstein. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39(4):783–791, 1985.

J. Felsenstein. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. 2004. URL https://evolution.genetics.washington.edu/phylip.html.

M. C. Fontaine, J. B. Pease, A. Steele, R. M. Waterhouse, D. E. Neafsey, I. V. Sharakhov, X. Jiang, A. B. Hall, F. Catteruccia, E. Kakani, S. N. Mitchell, Y.-C. Wu, H. A. Smith, R. R. Love, M. K. Lawniczak, M. A. Slotman, S. J. Emrich, M. W. Hahn, and N. J. Besansky. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science*, 347(6217):1258524, 2015. doi: 10.1126/science.1258524. URL https://www.science.org/doi/abs/10.1126/science.1258524.

T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proc Natl Acad Sci*, 101 (suppl_1):5228–5235, 2004.

D. M. Hillis and J. J. Bull. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Sist Biol*, 42(2):182–192, 1993.

M. T. Holder, J. Sukumaran, and P. O. Lewis. A justification for reporting the majority-rule consensus tree in Bayesian phylogenetics. *Sist Biol*, 57(5):814–821, 2008.

R. R. Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18:337–338, 2002.

M. La Rosa, A. Fiannaca, R. Rizzo, and A. Urso. Probabilistic topic modeling for the analysis and classification of genomic sequences. *BMC Bioinformatics*, 16(6):S2, 2015.

L. Liu, L. Yu, L. Kubatko, D. K. Pearl, and S. V. Edwards. Coalescent methods for estimating phylogenetic trees. *Mol Phylogenet Evol*, 53(1):320–328, 2009a.

L. Liu, L. Yu, D. K. Pearl, and S. V. Edwards. Estimating species phylogenies using coalescence times among sequences. *Syst Biol*, 58(5):468–477, 2009b.

L. Liu, L. Tang, W. Dong, S. Yao, and W. Zhou. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, 5(1):1608, 2016.

G. Marçais and C. Kingsford. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764–770, 2011.

B. Q. Minh, H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, and R. Lanfear. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol*, 37(5):1530–1534, 2020.

S. Mirarab, M. S. Bayzid, and T. Warnow. Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. *Syst Biol*, 65(3): 366–380, 2016.

L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, and B. Q. Minh. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*, 32(1):268–274, 2015.

T. H. Ogden and M. S. Rosenberg. Multiple sequence alignment accuracy and phylogenetic inference. *Syst Biol*, 55(2):314–328, 2006.

R. Řehůřek and P. Sojka. Software framework for topic modelling with large corpora. In *Proc LREC 2010 Workshop New Challenges NLP Frameworks*, pages 45–50. ELRA, 2010.

J. Ren, X. Bai, Y. Y. Lu, K. Tang, Y. Wang, G. Reinert, and F. Sun. Alignment-free sequence analysis and applications. *Annu Rev Biomed Data Sci*, 1(1):93–114, 2018.

A. Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.

J. Sukumaran and M. T. Holder. DendroPy: a Python library for phylogenetic computing. *Bioinformatics*, 26(12):1569–1571, 2010.

D. Swofford. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4., 2003.

S. Vinga and J. Almeida. Alignment-free sequence comparison—a review. *Bioinformatics*, 19(4):513–523, 2003.

D. Wen, Y. Yu, M. W. Hahn, and L. Nakhleh. Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. *Molecular Ecology*, 25(11):2361–2372, 2016. doi: https://doi.org/10.1111/mec.13544. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.13544.

K. M. Wong, M. A. Suchard, and J. P. Huelsenbeck. Alignment uncertainty and genomic analysis. *Science*, 319(5862):473–476, 2008.

C. Zhang, M. Rabiee, E. Sayyari, and S. Mirarab. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, 19(6):153, 2018.

A. Zielezinski, H. Z. Girgis, G. Bernard, C.-A. Leimeister, K. Tang, T. Dencker, A. K. Lau, S. Röhling, J. J. Choi, M. S. Waterman, M. Comin, S.-H. Kim, S. Vinga, J. S. Almeida, C. X. Chan, B. T. James, F. Sun, B. Morgenstern, and W. M. Karlowski. Benchmarking of alignment-free sequence comparison methods. *Genome Biol*, 20(1):144, 2019.

Electronic Supplement

Anopheles trees

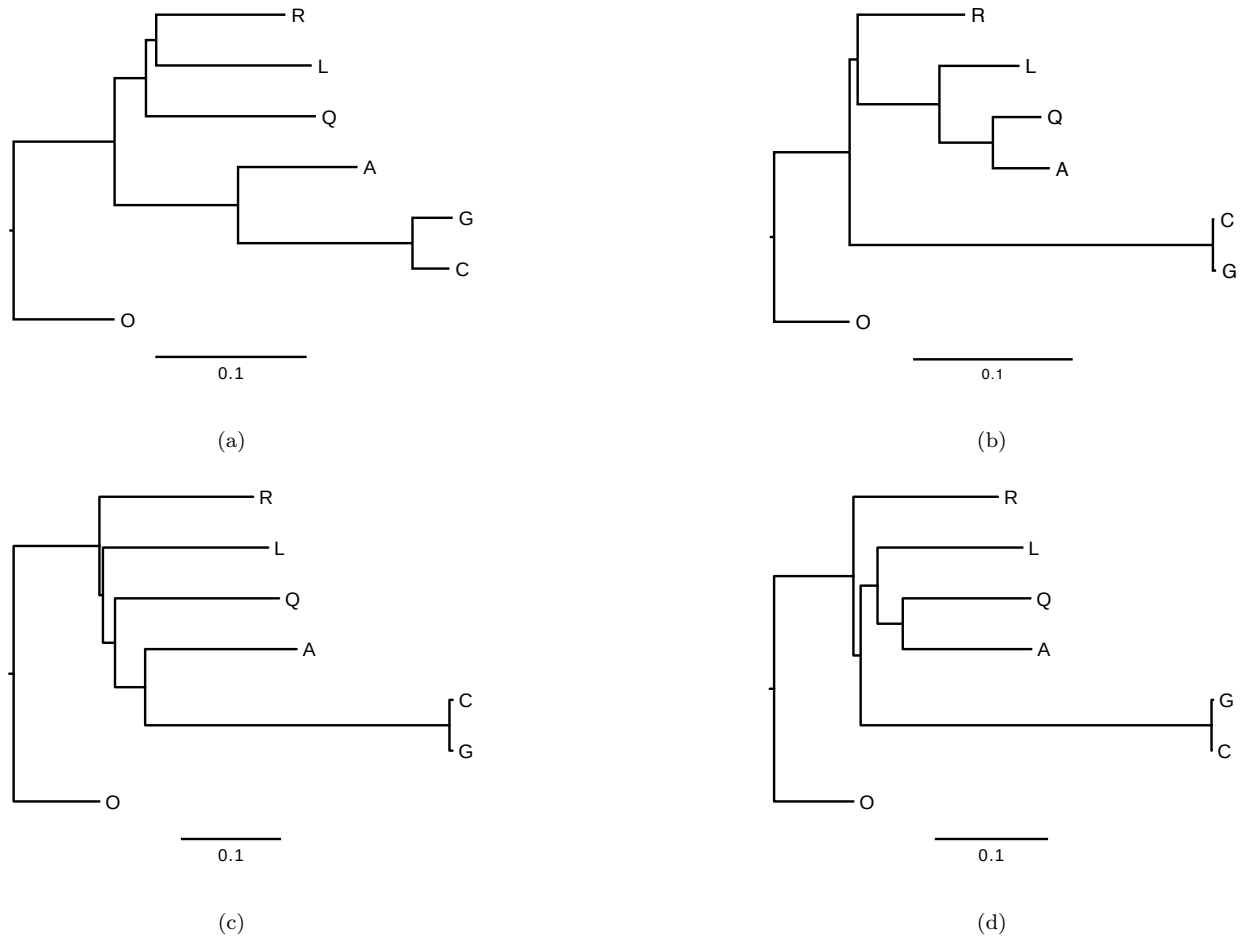

Figure S1. TopicContml analyses of 6 species in the *Anopheles gambiae* complex and an outgroup, using the data of Wen, D.; Yu, Y.; Hahn, M. W.; Nakhleh, L. [2016]. Data from: Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. Dryad. https://doi.org/10.5061/dryad.tn47c. *Anopheles gambiae* (G), *Anopheles coluzzii* (C), *Anopheles arabiensis* (A), *Anopheles quadriannulatus* (Q), *Anopheles merus* (R) and *Anopheles melas* (L). *Anopheles christyi* (O) is the outgroup to root the tree. (a) 67317 aligned autosomal loci; (b) 4157 aligned X-chromosome loci. (c) 67317 unaligned autosomal loci; (d) 4157 unaligned X-chromosome loci.