Software for Systematics and Evolution

Syst. Biol. XX(XX):1–12, 2025

https://doi.org/10.1093/sysbio/syaf015
Advance Access Publication February 25, 2025

Estimating Genome-Wide Phylogenies Using Probabilistic Topic Modeling

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

[1]*Department of Scientific Computing, Florida State University, 110 N Woodward Ave, Tallahassee, FL 32306, USA*
[2]*Department of Organismic and Evolutionary Biology and Museum of Comparative Zoology, Harvard University, 26 Oxford Street, Cambridge, MA 02138, USA*
**Correspondence to be sent to: Department of Scientific Computing, Florida State University, Tallahassee, FL 32306, USA; E-mail: mk16e@fsu.edu.*

Received 19 February 2024; reviews returned 01 February 2025; accepted 20 February 2025
Associate Editor: Matthew Hahn

Abstract.—Methods for rapidly inferring the evolutionary history of species or populations with genome-wide data are progressing, but computational constraints still limit our abilities in this area. We developed an alignment-free method to infer genome-wide phylogenies and implemented it in the Python package TopicContml. The method uses probabilistic topic modeling (specifically, Latent Dirichlet Allocation) 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. We evaluated the performance of TopicContml on simulated datasets with gaps and three biological datasets: 1) 14 DNA sequence loci from two Australian bird species distributed across nine populations, 2) 5162 loci from 80 mammal species, and 3) raw, unaligned, nonorthologous PacBio sequences from 12 bird species. We also assessed the uncertainty of the estimated relationships among clades using a bootstrap procedure. Our empirical results and simulated data suggest that our method is efficient and statistically robust. [Alignment-free; bootstrap; Contml; *k*-mers; LDA; multilocus phylogeny; NLP; topic modelling.]

Phylogenetic analysis traditionally relies on the alignment of orthologous sequence data, a process that can be challenging due to the complexity of genomic variations, difficulties in aligning noncoding regions, and the presence of highly divergent sequences. Over the past two decades, alignment-free approaches based on shared properties of subsequences of defined length *k* (*k*-mers or *k*-grams) (Deschavanne et al., 1999; Chapus et al., 2005; Shedlock et al., 2007; Marçais and Kingsford, 2011) have been developed to compare sequences and genomes. Alignment-free methods for evolutionary analysis have been reviewed (Vinga and Almeida, 2003; Zielezinski et al., 2019) and their robustness 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; Balaban et al., 2022; Van Etten et al., 2023). Several studies have shown that alignment artifacts can significantly impact tree topology (Ogden and Rosenberg, 2006; Wong et al., 2008; Du et al., 2019). Alignment becomes problematic with comparisons of large genomes, complex genomic variations, challenges in aligning noncoding regions, difficulties presented by highly divergent sequences, and the time-consuming nature of aligning large datasets. Alignment-free approaches offer a promising alternative to address these weaknesses of alignment-based methods (Ren et al., 2018).

Probabilistic topic modeling (Blei et al., 2003) is a statistical approach aiming to identify major "themes," "connections," or "topics" among themes in documents and other large collections of text. The approach 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 a natural grouping (topics) of documents based on the content from a large number of scientific documents, called a corpus. Latent Dirichlet Allocation (LDA) is a popular technique in topic modeling within the context of unsupervised machine learning, introduced by Blei et al. (2003). The goal is to uncover these topics by analyzing the words in the documents and essentially "learning" the structure of the data. The method assumes that documents consist of latent topics, each represented by a distribution of words. LDA has also been a focus of 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 the analysis of single-cell expression or epigenetic data (duVerle et al., 2016; Dey et al., 2017). Here, we present a novel computational approach using probabilistic topic modeling to infer evolutionary relationships among individuals from different populations

or species. This method works with multilocus data, including unaligned or aligned DNA sequences and unassembled raw sequencing reads. We will use the term species tree for these trees, whether derived from single individual sequences or individuals grouped into populations or species.

Genome-wide datasets (sequences of whole genomes or multiple genes per species) are becoming increasingly prominent in inferring the evolutionary history of closely related species. Traditional approaches to multilocus phylogenetics, such as concatenation methods (Gatesy and Baker, 2005), and approaches that are consistent under the multispecies coalescent (de Queiroz and Gatesy, 2007; Liu et al., 2009; Chifman and Kubatko, 2014; Zhang et al., 2018; Mirarab et al., 2016), have advanced the field significantly. However, these methods often rely on high-quality alignments, which can be computationally expensive and error-prone when dealing with large or complex datasets. TopicContml offers a scalable and efficient alternative to traditional approaches, addressing their limitations and enabling the analysis of diverse and complex genomic datasets without relying on sequence alignment.

Here, we outline the architecture of TopicContml and demonstrate its application using simulated data and three empirical datasets: 1) a small orthologous dataset of individuals from various locations of two parapatric bird species, consisting of 14 loci, along with a comparison to SVDquartets (Chifman and Kubatko, 2014) and the alignment-free method Mash (Ondov et al., 2016), including a discussion of bootstrap support; 2) an orthologous dataset of mammalian species, consisting of 5162 loci across 90 vertebrate species; and 3) a 12-species bird dataset with unaligned, nonorthologous PacBio raw sequencing reads.

Materials and Methods

TopicContml Software

TopicContml is a Python package based on a two-phase pipeline: 1) The mulitlocus or genome-wide sequences are fragmented into k-mers; these k-mers are then used to learn a probabilistic topic model and extract the topic frequencies of these k-mers using the LDA model for each locus (Fig. 1 left). For a data analysis with multiple individuals from the same species or population, we set the option `--merging n` to merge individual labels that start with the same n letters into groups; otherwise, we assume that each sequence is an individual. 2) These topic frequencies from multiple loci are then used to estimate a phylogeny with Continuous Characters Maximum Likelihood (Contml) (part of PHYLIP; Felsenstein, 1981, 2004) (Fig. 1 right).

K-mer decomposition.—In NLP, large text datasets (corpora) are broken down into smaller units such as documents, which are further divided into words or sentences, referred to as tokens. Similarly, in bioinformatics, datasets consisting of multiple genomes or multilocus DNA sequences can be broken down into individual genomes or groups of multilocus sequences associated with an individual. These sequences are then decomposed into k-mers—substrings of length k representing short DNA or amino acid sequences. We decompose the DNA sequences into nonoverlapping k-mers, as shown in Figure 1, since overlapping k-mers require more memory and computation time without yielding significantly better results. The program estimates an optimal k-mer length based on the probability of observing a given k-mer in a document (sequence) at each locus, with options for user adjustments.

The probability of a given k-mer K appearing in a random genome X of size n is $P(K \in X) = 1 - \left(1 - |\Sigma|^{-k}\right)^n$, where $\Sigma = \{A, C, G, T\}$, without loss of generality. Given a document size n and the desired probability q of observing a random k-mer, the value of k that minimizes the probability of observing a random k-mer can be computed as (Fofanov et al., 2004; Ondov et al., 2016)

$$\hat{k} = \lceil log_{|\Sigma|}(n(1-q)/q) \rceil.$$

TopicContml calculates $\hat{k}$ for each document in a locus based on this. We have found that $k = 20$ and $k = 8$ give accurate estimates in most cases for large (e.g., 1,000,000 bp) and small sequences, respectively. The program also allows users to choose different k-mer configurations, such as a combination of k-mer lengths or a single fixed length, based on their analysis needs.

Resolving ambiguities and missing data.—The current version of TopicContml retains all IUPAC codes as they are, except for "N," "?," "-," and other ambiguous characters, which can be filtered out prior to analysis. We assume that such ambiguity codes are rare and do not strongly affect the results.

Topic modeling.—Given a collection of D documents and a number of M topics, topic modeling discovers 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 is a generative probabilistic model used to uncover hidden topics within a collection of documents, referred to as a corpus. Given a corpus with D documents, let N be the number of words in a specific document $d \in D$. For each word $w_{d,n}$ (the nth word in the dth document), $z_{d,n}$ denotes the associated topic. The distribution of topics for document d, represented by θ_d, is drawn from a Dirichlet distribution, $\theta_d \sim Dir(\alpha)$, where $\alpha > 0$ is the parameter vector. Similarly, the distribution of words for each topic m, denoted by β_m, is also drawn from a Dirichlet distribution, $\beta_m \sim Dir(\eta)$, with parameter vector $\eta > 0$. In this model, the only observed variables are the words w in the documents, while the topics z, the topic

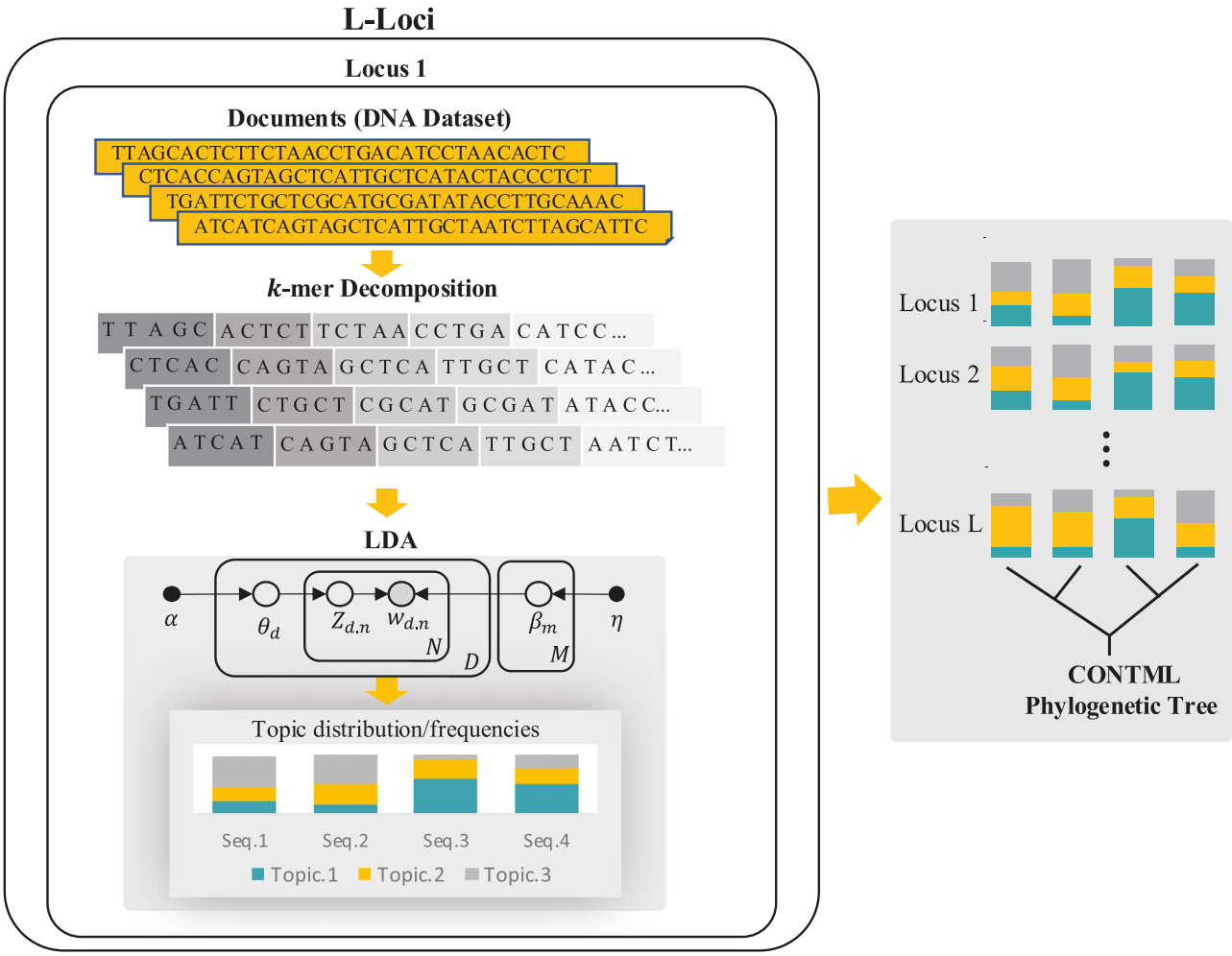

FIGURE 1. TopicContml workflow to generate topic frequencies and the corresponding phylogeny.

distributions θ for all documents, and the word distributions β for each topic are all latent variables. 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 this joint distribution, one can compute the posterior distribution of the unknown model parameters, $p(\theta,z,\beta|w) = \frac{p(\theta,z,\beta,w)}{p(w)}$, using expectation-propagation (Minka and Lafferty, 2012) or other maximization methods.

For each locus, TopicContml first estimates the topic frequencies for every document, θ, using the Python package Gensim (Řehůřek and Sojka, 2010) (Fig. 1). The sequences (documents) in each locus are decomposed into k-mers (words). During preprocessing, LDA filters out certain words, primarily those with low frequency, to improve topic coherence and reduce noise in the learned distributions (see Supplementary Fig. S8). The process begins by randomly assigning a distribution of topics to each document and a distribution of words to each topic, with these distributions being governed by Dirichlet priors. During the training phase, Gensim's implementation of LDA iteratively refines these distributions using maximization methods. This training involves updating two key parameters: the distribution of topics within each document and the distribution of words within each topic. The model trains by analyzing patterns of word co-occurrence across the documents, assigning words to topics in a way that maximizes the likelihood of the observed data. As the iterations progress, the model converges to a stable set of topics.

Determining the optimal number of topics.— Selecting the optimal number of topics in LDA modeling is important for generating interpretable results. A widely used method for this involves evaluating topic coherence, which reflects the semantic similarity among top words within each topic—higher coherence scores generally

correlate with more interpretable topics (Röder et al., 2015). Several algorithms are available for calculating coherence scores. In this study, we use the "u_{mass}" coherence measure in Gensim (Řehůřek and Sojka, 2010), analyzing each locus individually to identify the best topic number for each (see Supplementary Fig. S3). However, because coherence analysis involves testing multiple models with varying topic counts, it can be time-intensive, especially with large datasets. To address this, TopicContml also allows users to specify a fixed number of topics, which, in our case, proved to be a practical alternative without compromising the interpretability or consistency of results. Although selecting an optimal topic count can enhance detail in some studies, our findings show that a fixed topic number performs well and offers efficiency for large-scale analyses. All our analyzed datasets are based on a fixed value of five topics.

Visualizing topics and associated terms.—To visualize the topics and associated terms (*k*-mers), we use the package pyLDAvis (Sievert and Shirley, 2014), an interactive visualization tool. TopicContml allows users to generate an HTML file containing the pyLDAvis output for each locus, saving them in a designated folder within the directory (see Supplementary Fig. S2). This enables detailed examination and extraction of information from the fitted LDA model, enhancing our ability to interpret the underlying topic structures.

Contml (Continuous Characters Maximum Likelihood method).—The results of the LDA analysis are the topic frequencies for each document, which are then evaluated using Contml to estimate a phylogeny from frequency data using the restricted maximum likelihood method (Felsenstein, 1973) based on the Brownian motion model for allele frequencies (Cavalli-Sforza and Edwards, 1967). The primary assumption of Contml is that each character (each topic in our case) evolves independently according to a Brownian motion process and that character state changes since divergence are independent of each other, which means that the net change after t units of time is normally distributed with zero mean and variance ct (the same constant c for all characters).

Multilocus bootstrapping.—To assess the statistical confidence of the inferred phylogenies, we use bootstrapping (Efron, 1979). Given a multiple sequence alignment, the bootstrap method involves resampling the original dataset with the 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 a majority-rule consensus tree (Hillis and Bull, 1993; Efron et al., 1996; Holder et al., 2008). TopicContml implements bootstrapping strategies for both aligned and unaligned sequences. It generates a majority-rule consensus tree from the bootstrap replicates using SumTrees in DendroPy (Sukumaran and Holder, 2010).

Datasets

We evaluate the accuracy of TopicContml for simulated data and multiple real biological data. The simulated data were used to explore the effects of the number of loci and the accuracy of recovering the true topology from data sets consisting of 7 and 14 species. We used three biological datasets with very different features: 1) a 14-locus dataset from 2 parapatric closely related bird species separated into 9 populations, to evaluate the accuracy of estimation, bootstrapping, and to compare accuracy with SVDQuartets (Chifman and Kubatko, 2014) and the alignment-free approach Mash (Ondov et al., 2016); 2) a vertebrate dataset focusing on mammals with 90 species and 5162 loci to evaluate the effect of missing data and of aligned versus nonaligned orthologous loci; 3) a dataset of raw PacBio sequences of 12 bird species, each containing 100,000 reads; these sequences were neither orthologous nor aligned and can be construed as having been sampled randomly and potentially overlapping from their constituent genomes.

Simulated datasets.—We evaluated two sets of simulations: one with moderate and one with many insertions and deletions. We used the software Dawg (Cartwright, 2005) to simulate aligned sequences with indels/deletions on a 7-species and a 14-species tree (Fig. 2). The moderate scenario inserts indels at a rate of 0.02 per site and deletes sites with a rate of 0.02 per site. The more extreme simulation used an insertion/deletion rate of 0.2 per site. An example of two individual sequences for each indel/deletion scenario is shown in the electronic supplement (Supplementary Fig. S1). We simulated datasets of 1, 2, 5, 10, 20, 50, 100, 200, 500, and 1000 loci; each locus was between 800 and 2000 bp, dependent on indels/deletions. These simulated datasets were then analyzed with gaps included (aligned), with the gaps excised (unaligned), and with *k*-mers that include gaps removed (no gap-kmer). We fixed the number of topics for all simulation analyses to 5 and estimated the best *k*-mer size from the data; it turned to be 9 for all datasets. Each scenario was run 100 times. The resulting trees were compared with the true trees using unweighted and weighted Robinson–Foulds (RF) distances.

Empirical datasets.—We evaluated three biological datasets:

Two parapatric closely related bird species: The data contain multiple individuals of the Australian brown treecreeper (*Climacteris picumnus*) and black-tailed tree creeper (*Climacteris melanurus*) (Edwards et al., 2022, 2023). DNA sequences consisted of 14 loci and 9 different geographic locations. For each locus, sequence

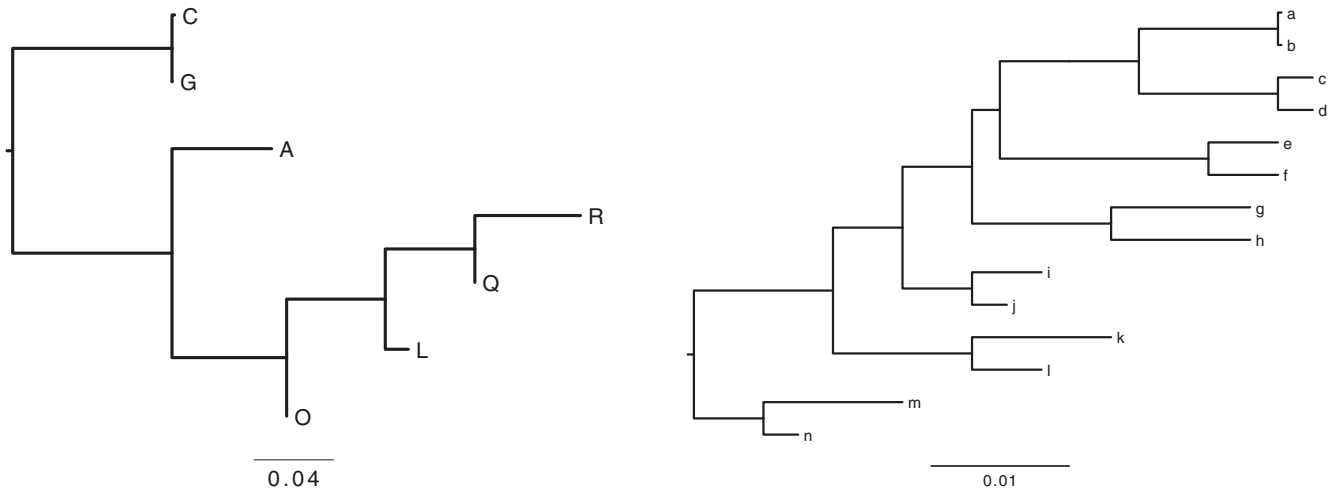

FIGURE 2. Phylogenies of 7 and 14 species used for the simulation of sequence data with deletions and indels (gaps). In Tables 1 and 2, these are referenced as 7-tip and 14-tip trees.

length varied from 288 to 418 base pairs, and the number of aligned sequences per locus ranged from 78 to 92. For the evaluations of unaligned data, we removed all gaps in each sequence. We applied LDA to each locus across all nine locations and extracted the topic frequencies. These topic frequencies were used in CONTML to generate the population tree, evaluate bootstrap support, and compare with another approach.

Mammal dataset: The second dataset includes 90 vertebrate species focusing on mammals with 5162 loci (Wu et al., 2018) and was analyzed by Liu et al. (2017) using concatenated maximum likelihood and coalescent approaches. We analyzed this dataset using TOPICCONTML under four conditions: 1) excluding *k*-mers containing gaps ("-") or unspecified nucleotides ("N"), 2) removing alignment columns with gaps before excluding *k*-mers with "N," 3) removing all gaps from sequences before excluding *k*-mers with "N," and 4) using aligned sequences while retaining "N" characters. We then used the RF distance to compare each phylogram with the maximum-likelihood tree and random trees.

PACBIO dataset: A total of 6.5 GB of raw sequencing reads of 12 bird species generated by the PACBIO HIFI sequencing method. The 12 birds cover most of the depth of the avian tree, including the two deepest branches; Paleognathae and Neognathae Species were chosen so as to broadly sample the tree for species belonging to lineages whose higher relationships are fairly stable, but also to include unambiguously close relatives, so that we could test the ability of our method to recover close relatives (Jarvis et al., 2014; Oliveros et al., 2019). We subsampled 100,000 reads from each species; average read length per species varied from 9.3 kb in the tinamou Crypterellus tataupa to 18.8 kb in chicken.

RESULTS

Analysis of Simulated Datasets

The simulations based on the moderate insertion/deletion scheme, shown in Table 1, demonstrate that the recovery of trees close to the true tree improves with the number of loci for all treatments and also for both tree topologies. The "Aligned" treatment for both tree topologies works well with more than 50 loci. The "Not aligned" treatment worked better for the 14-species tree than for the 7 species tree, but we used only a small number of replicates ($n = 100$). Still, accurate recovery of the true tree was almost as high as with the "Aligned" treatment. The "No gap-kmer" treatment fairs as well as the "Not aligned" treatment. Weighted RF distances (wRF) show trends consistent with the percentage of trees recovered that either match the true tree or are within two distance units of it. Notably, the 'No gap-kmer' treatment is closer to the true tree than the 'Not aligned' treatment, which performs comparably to the 'Aligned' treatment.

The simulations with extreme insertion/deletion strategy (as shown in Table 2) are markedly different for the 'Not Aligned' treatment of the 14 tip trees: even with 1000 loci none of the estimated trees were close to the true tree; 'Aligned' and 'No gap-kmer' faired similar to the moderate indel/deletion scheme. For the 7-species tree all treatments are close to the true tree when the number of loci is large. Overall the wRF values are all higher than those for the moderate indel/deletion scenario, and the "No gap-kmer" delivers similar values as the "Aligned."

Analysis of Empirical Datasets

Multilocus species tree from closely related Australian birds.—For the treecreeper dataset, we tokenized each

TABLE 1. Simulated data with low indel/deletion frequency and accuracy of phylogenetic reconstruction with TopicContml: data were simulated using Dawg (Cartwright, 2005), each locus has around 1000 bp with gaps using a low number of gap scenario (parameter: insertion/deletion probability 0.02/site, average indel length 12, topics=5, *k*-mers were estimated from the data (all estimated *k*-mers were 9 base pairs); 'Aligned' results are based on simulated data with gaps, 'Not Aligned' had all gaps removed in the simulated data, "No gap-kmer" had all *k*-mers containing gaps removed before LDA. "C" (Close) marks the frequency of topologies that are either the same as the true topology or not more than 2 rearrangments apart; "wRF" is the average weighted RF distance from the true tree. Each result is based on 100 simulations.

Loci	7 Species						14 Species					
	Aligned		Not Aligned		No gap-kmer		Aligned		Not Aligned		No gap-kmer	
	C	wRF	C	wRF	C	wRF	C	wRF	C	wRF	C	wRF
1	0.33	1.30	0.27	1.26	0.39	1.30	0.00	1.84	0.00	1.68	0.01	1.81
2	0.61	1.21	0.35	1.25	0.54	1.22	0.02	2.01	0.00	2.07	0.03	1.95
5	0.64	1.20	0.53	1.22	0.64	1.20	0.24	2.01	0.13	2.17	0.27	1.96
10	0.70	1.18	0.47	1.19	0.69	1.18	0.35	2.03	0.31	2.14	0.40	2.00
20	0.78	1.17	0.67	1.20	0.78	1.18	0.64	2.00	0.74	2.14	0.58	2.00
50	0.85	1.17	0.76	1.18	0.70	1.19	0.83	2.01	0.90	2.14	0.81	1.99
100	0.93	1.17	0.74	1.20	0.87	1.19	0.93	1.98	0.96	2.14	0.92	1.97
200	0.97	1.17	0.87	1.19	0.95	1.17	0.95	1.98	1.00	2.13	0.96	1.97
500	0.99	1.17	0.88	1.20	0.99	1.17	0.99	1.97	1.00	2.13	1.00	1.96
1000	1.00	1.17	0.94	1.19	1.00	1.18	1.00	1.97	1.00	2.13	1.00	1.95

TABLE 2. Simulated data with high indel/deletion frequency and accuracy of phylogenetic reconstruction with TopicContml: data were simulated using Dawg (Cartwright, 2005), each locus has around 1000 bp with gaps using a high number of gap scenario (parameter: insertion/deletion probability 0.2/site, average indel length 12, topics=5; *k*-mers were estimated from the data (all estimated *k*-mers were 9 base pairs); "Aligned" results are based on simulated data with gaps, "Not Aligned" had all gaps removed in the simulated data, "No gap-kmer" had all *k*-mers containing gaps removed before LDA. "C" (Close) marks the frequency of topologies that are either the same as the true topology or not more than 2 rearrangments apart; "wRF" is the average weighted RF distance from the true tree. Each result is based on 100 simulations.

Loci	7 Species						14 Species					
	Aligned		Not Aligned		No gap-kmer		Aligned		Not Aligned		No gap-kmer	
	C	wRF	C	wRF	C	wRF	C	wRF	C	wRF	C	wRF
1	0.40	1.25	0.18	1.30	0.22	1.28	0.00	1.86	0.00	2.06	0.00	1.85
2	0.47	1.17	0.23	1.22	0.48	1.22	0.07	2.03	0.00	2.96	0.03	1.98
5	0.60	1.19	0.30	1.23	0.56	1.18	0.16	2.08	0.00	3.31	0.17	2.00
10	0.61	1.19	0.42	1.21	0.57	1.20	0.29	2.09	0.00	3.65	0.36	2.03
20	0.62	1.20	0.57	1.23	0.57	1.22	0.58	2.11	0.00	3.70	0.63	2.02
50	0.59	1.22	0.75	1.20	0.45	1.22	0.84	2.10	0.02	3.78	0.75	2.02
100	0.66	1.21	0.76	1.21	0.58	1.22	0.93	2.09	0.01	3.78	0.93	2.00
200	0.74	1.22	0.86	1.21	0.56	1.23	0.97	2.07	0.02	3.82	0.97	1.99
500	0.88	1.20	0.93	1.22	0.55	1.22	0.99	2.06	0.01	3.83	1.00	1.98
1000	0.82	1.21	1.00	1.22	0.59	1.22	1.00	2.05	0.02	3.81	1.00	1.98

DNA sequence at every locus using *k*-mer representation with a *k* value of 8 (as estimated by TopicContml), employing nonoverlapping tokens. Individuals from the same location were merged, and LDA was applied to the corpus to generate topic frequencies for each locus for each of the nine populations. We used five topics in our analysis. These multilocus topic frequencies were then used to construct a maximum-likelihood tree with Contml in the PHYLIP package. We did bootstrapping, and Figure 3b shows the majority-rule consensus tree generated by TopicContml for unaligned data, with bootstrap support values derived from 1000 replicates. Inspection of the clades reveals that our tree recovers the expected geographic relationship within each species, and the locations are separated by species, which in turn are separated by the Carpentarian barrier in Australia (Cracraft, 1986; Edwards et al., 2023).

We compared the performance of TopicContml with SVDquartets (Chifman and Kubatko, 2014) implemented in Paup* (Swofford, 2003) (SVDquartets +Paup*), as shown in Figure 3c. There are notable differences between the results of SVDquartets and TopicContml. When comparing these results to the mapped locations in Figure 3a, we observe that our bootstrap tree from TopicContml recovers the relationships equally well or better than SVDquartets. Both methods encounter challenges in resolving certain population splits but confidently separate the two species. TopicContml support values recover, in general, the geographic pattern of the locations well.

We also compared our phylogenetic tree with one generated using the alignment-free approach Mash (Ondov et al., 2016), which estimates evolutionary distances between nucleotide sequences. For the Mash input, sequences from different loci for each individual were concatenated, with missing data filled by gaps. The sequences from individuals at the same location were then merged to create nine Fasta files, representing the nine populations in Australia. The pairwise distance matrix generated by Mash was used to construct a Neighbor-Joining tree using the PHYLIP package (Felsenstein, 2004). Figure 3d shows the phylogenetic

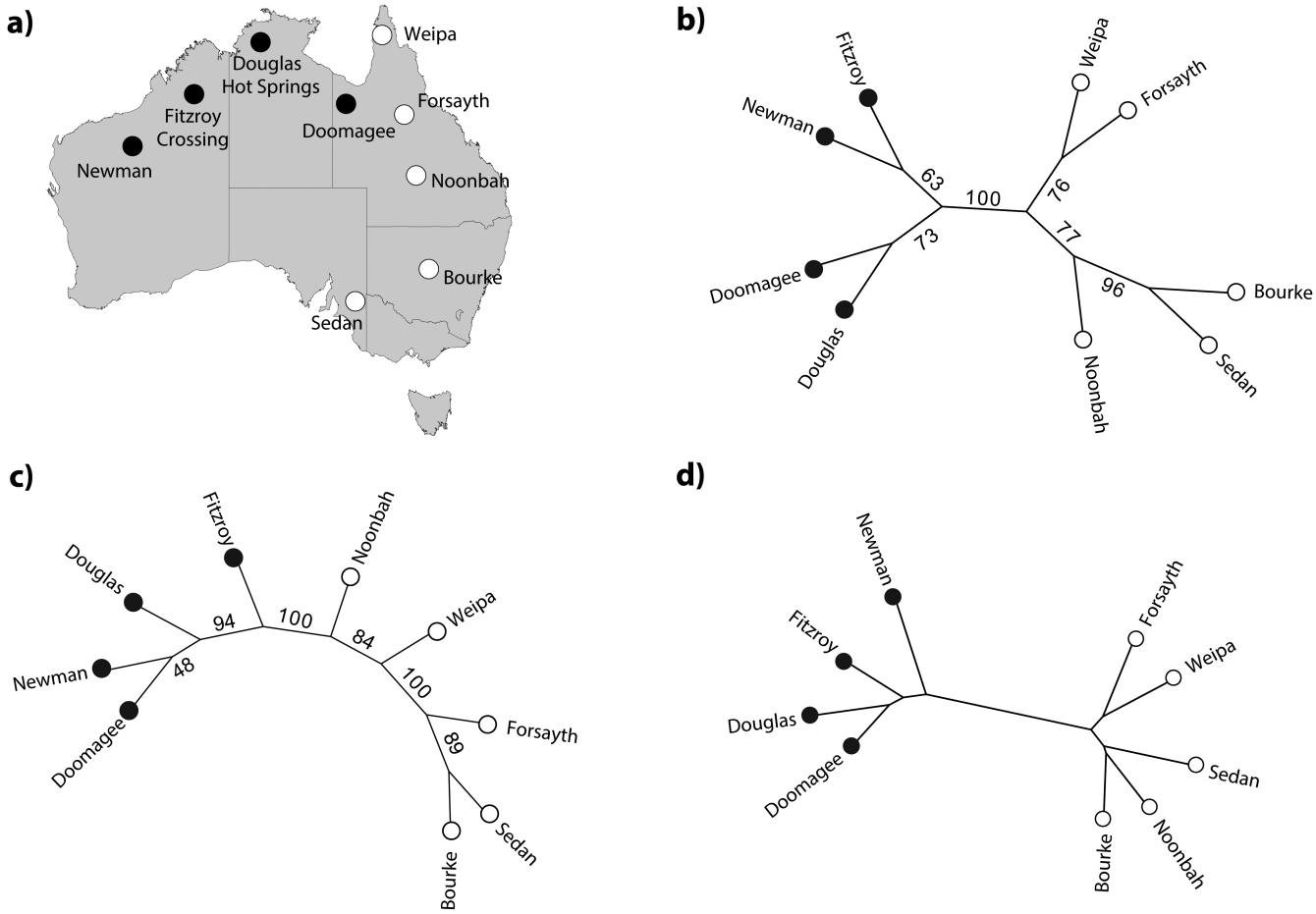

FIGURE 3. Relationship tree of nine populations of two Australian treecreeper species reconstructed. (a) The map of Australia shows the locations; the black disks mark *Climacteris melanurus*, and the white disks mark *Climacteris picumnus* (Edwards et al., 2023). (b) The majority-rule consensus tree of unaligned data by TopicContml. For each bootstrap analysis, 1000 replicates were used. Values in the graphs are % support. (c) The majority-consensus tree of aligned data analyzed by SVDQuartets +Paup*. (d) The phylogeny constructed by Mash.

tree generated by Mash. When comparing our tree to the one generated by Mash, we observe that both trees depict similar relationships among species.

Multilocus species tree from genome-wide mammal dataset.— We analyzed a mammal data with 90 species and 5162 loci . The dataset consisted of nucleotide characters from the set "ACGT-N." We analyzed the mammal dataset under various conditions and compared the resulting trees to the maximum likelihood tree derived from 4388 loci of 90 vertebrate species, as reported by Liu et al. (2017). The comparisons were visualized using tanglegrams (Revell, 2024). In the first analysis, after generating *k*-mers, we excluded any *k*-mers containing either "-" or "N." The resulting phylogenetic tree (Fig. 4) was 60 steps away from the maximum likelihood tree, as measured by the RF distance. In the second analysis, we removed all alignment columns containing gaps, then excluded *k*-mers with "N." The resulting tree (Supplementary Fig. S4) was 64 steps from the reference tree. This resulted in a tree generated using 1719 loci because 3443 of the 5162 loci did not contain any data after the removal. In the third analysis, we removed all gaps from each sequence before excluding *k*-mers containing "N." This approach did not improve the tree (Supplementary Fig. S5), which was 90 steps from the reference tree. Finally, in the fourth analysis, we generated a tree using aligned sequences while retaining "N" characters. This tree (Supplementary Fig. S6) was 80 steps from the maximum likelihood tree.

The RF distances between 1,000,000 random trees and the maximum likelihood tree reveal a minimum distance of 168, a mean distance of 173.5, and a maximal distance of 174. The topic modeling trees are therefore considerably closer to the maximum likelihood tree than random trees (Supplementary Fig. S7), confirming that our topic modeling approach recovers phylogenetic signal. The tanglegrams (Fig. 4; Supplementary Figs. S4–S6) also confirm that our tree and the maximum likelihood tree are fairly similar despite a seemingly large RF distance, especially when considering that many of the branches in the mammal tree are very short (Foley et al., 2023).

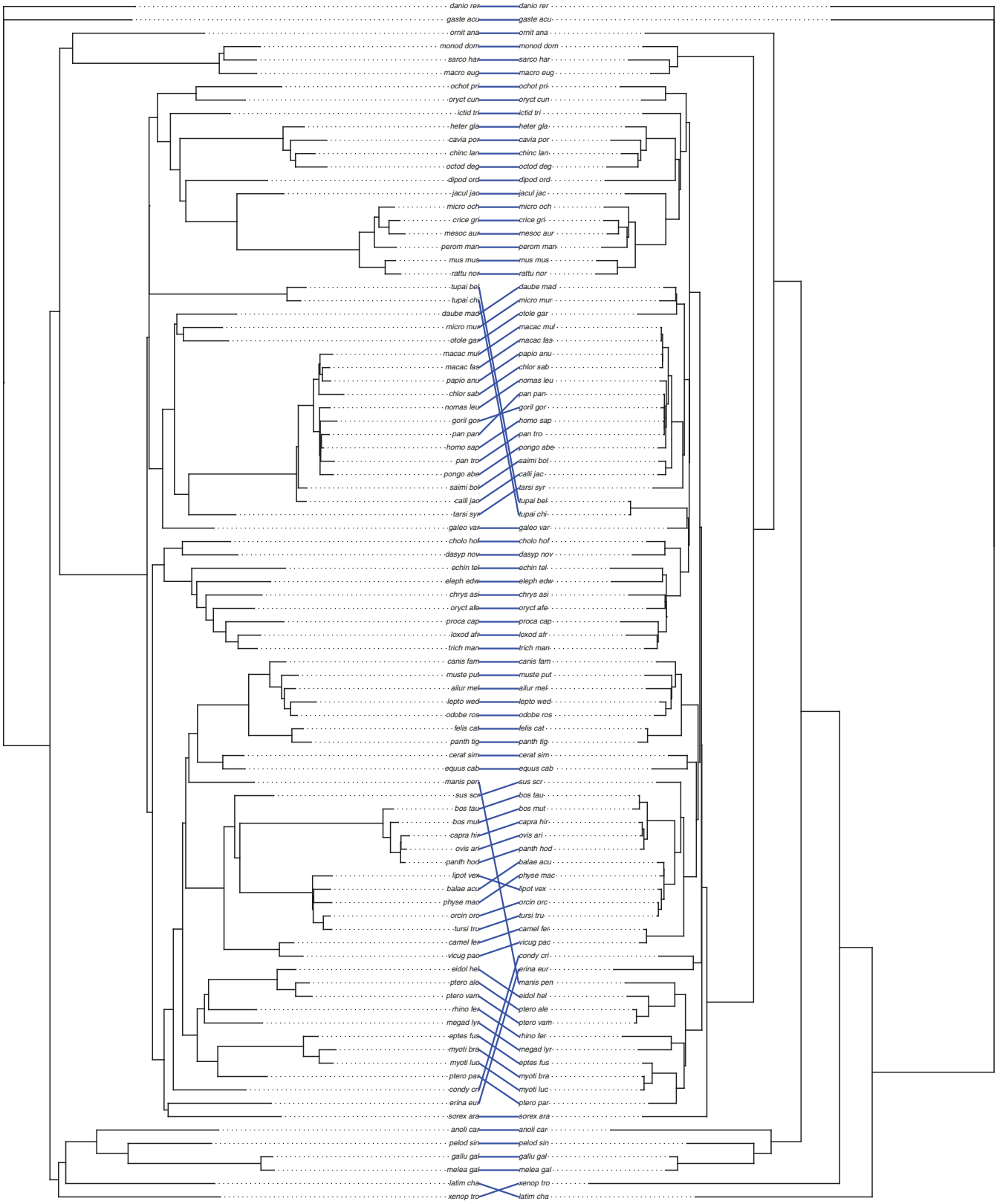

FIGURE 4. Tanglegram of mammal dataset comparing the TOPICCONTML tree (left), generated by excluding k-mers containing "-" or "N," with the maximum likelihood tree from Liu et al. (2017) (right). The alphabetical list of the species names in the tree is in Supplementary Table S1.

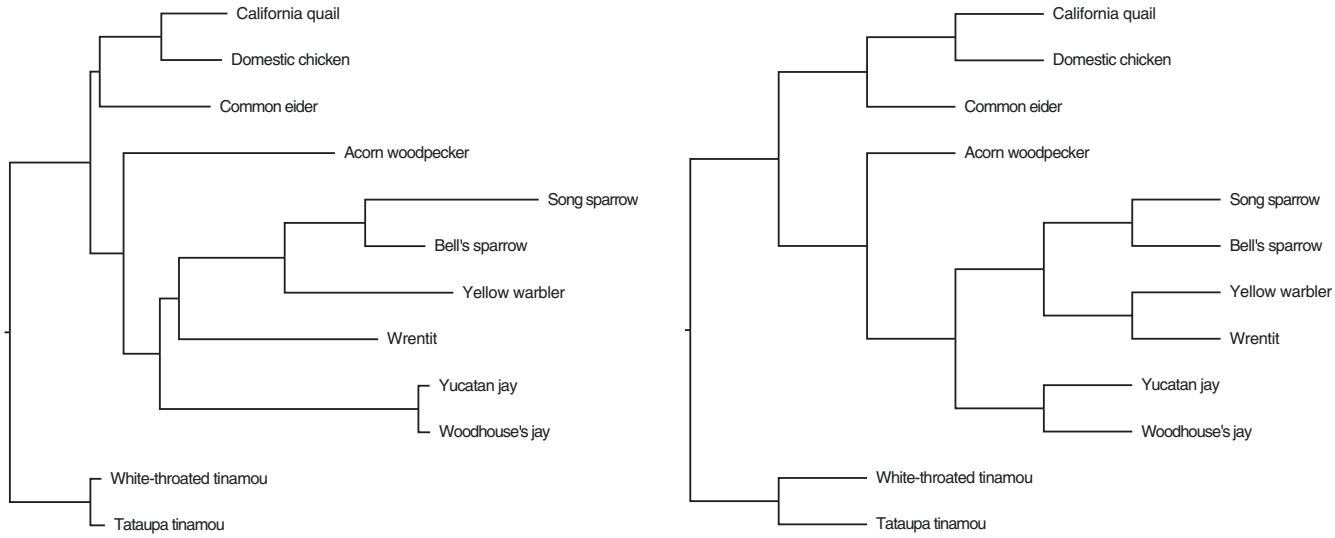

FIGURE 5. The phylogeny generated by TopicContml (left side) compared with the reference tree (right side).

Multilocus species tree from raw unassembled PacBio sequence reads.—This dataset consists of FASTA files containing 100,000 reads from each of 12 species. PacBio reads were counted and parsed with Seqkit stats (Shen et al., 2016) and seqtk sample (Li, 2013). To optimize computational efficiency, we concatenated every 1000 reads into single sequences, resulting in 100 loci per species. This approach balances the need to manage sequence length for LDA while reducing the total number of loci, thereby enhancing computational performance. Figure 5 (left) displays the tree generated by TopicContml using a nonoverlapping *k*-mer length of 20, with an RF distance of two from the reference tree, as shown in Figure 5 (right). The reference tree was drawn from Cracraft (1988), Oliveros et al. (2019), and Wu et al. (2024). The discrepancy in relationships of the Wrentit and Yellow warbler is uncertain, because relationships in this portion of the passerine tree are certainly not definitive (Oliveros et al., 2019).

Given the large size of each document, we performed our first analysis using a longer *k*-mer length (*k*-mer length of 20 as we discussed in Section *K*-mer decomposition). To assess the effects of different *k*-mer ranges, we experimented with lengths from 8 to 42. We found that optimal results were achieved with *k*-mers in the range of 18–25, as demonstrated in Supplementary Figure S9.

Discussion

This study integrates *k*-mers and probabilistic topic modeling to perform phylogenetic analysis on unaligned or aligned multilocus sequence data, as well as on unassembled raw sequencing reads. The Python code TopicContml offers an efficient workflow to reconstruct evolutionary relationships potentially without prior sequence alignment. TopicContml contains a two-phase analysis. First, *k*-mers are extracted from DNA sequences, and LDA uses these *k*-mers to establish how probable an individual's set of *k*-mers fits an arbitrary number of topics. For each locus and each individual, we generate a vector of assignment frequencies for a predefined set of topics. This step can be parallelized among loci. The LDA runtime depends on the length of the sequences (documents) and the number of loci. In the second step, the topic frequencies are used as input for Contml to construct a phylogenetic tree. The Contml evaluation time is influenced by the number of tips and the number of loci, as the input matrix for Contml is structured as number of tips × (number of loci × (number of topics - 1)). Although the current version of Contml does not support parallelization, implementing this capability could significantly improve its runtime, particularly for analyses involving many species. As shown in Supplementary Table S3, the Australian bird dataset completed in seconds for both LDA and Contml. The mammal dataset, however, took longer for LDA due to its large number of loci (5,162), despite relatively short sequence lengths, and even longer for Contml due to the large input size. For the PacBio dataset, the LDA runtime was extended by the significantly longer sequence lengths, but the Contml step completed within seconds.

The simulated data experiments reflect the influence of gap handling on phylogenetic reconstruction accuracy under different insertion/deletion (indel) rates and tree complexities. The simulation protocol produced aligned data with low and high gap numbers. The aligned data produced the most accurate phylogenetic inference; for both moderate and high indel frequencies. The observed decline in accuracy for the "Not Aligned" group, especially in the high-indel 14-species tree, stems

from the loss of indel information and, because gaps are removed entirely, the formation of new *k*-mer that do not necessarily coincide with the phylogenetic signal. The "Not Aligned" group faired well with the moderate indel scheme because only few *k*-mers were affected. The "No gap-kmer" group's results, which approximate those of the "Aligned" group under moderate indel scenarios, suggest that excluding only gap-containing *k*-mers strikes a balance between noise reduction and information preservation. This selective approach retains enough informative content while minimizing the alignment artifacts that become problematic in high-indel scenarios.

In contrast, the analysis of the real datasets was more complex, as demonstrated using the mammal dataset to evaluate the effect of structuring sequence data into *k*-mers. The RF distances to the maximum likelihood tree (Liu et al., 2017) reveal that increasing the number of loci does not always improve phylogenetic accuracy. The mammal dataset analysis highlights the robustness of TopicContml in recovering phylogenetic signal under various treatments of gaps and missing data. The most accurate tree (RF distance of 60) was achieved by excluding *k*-mers containing gaps or "N," underscoring the importance of targeted ambiguity removal. Removing entire alignment columns with gaps reduced the number of loci but still produced a comparable tree (RF distance of 64). Conversely, removing all gaps from sequences resulted in the least accurate tree (RF distance of 90), likely due to the loss of biologically informative gap signals or formation of phylogenetically noninformative *k*-mers created at the gap boundaries. Retaining "N" characters in aligned sequences yielded an intermediate result (RF distance of 80), showing that while ambiguity introduces noise, key phylogenetic relationships are still preserved. Notably, all TopicContml trees were substantially closer to the reference tree than random trees.

Tree uncertainty is commonly assessed through bootstrap analysis, which poses challenges for unaligned datasets as it requires bootstrapping at the *k*-mer level rather than the sequence level. For the treecreeper dataset, bootstrap analysis with TopicContml demonstrates its robustness in recovering phylogenetic relationships from unaligned data. The majority-rule consensus tree effectively separates the two species and accurately captures geographic patterns, including the division across the Carpentarian barrier. Compared with the alignment-based SVDquartets, TopicContml achieves equal or better precision in recovering geographic relationships. Furthermore, the comparison with the alignment-free method Mash confirms that TopicContml effectively captures key phylogenetic relationships while working with unaligned data.

Our analyses of the PacBio dataset shows substantial promise deriving phylogenetic signal from unaligned long-read sequences and demonstrates the potential of TopicContml for alignment-free phylogenetic reconstruction. Despite the complexity of raw, unassembled reads, TopicContml produced trees closely matching a reference phylogeny, showing its capacity to infer evolutionary relationships directly from complex, heterogeneous data. An important goal of the future is to determine what components of unaligned genomic data—tranposable elements, satellite sequences, or other common components of genomes—are driving these positive results. Although the LDA step for this dataset required additional processing time due to long-read lengths, Contml efficiently completed tree inference. These results suggest that TopicContml offers a promising approach for handling high-throughput phylogenetic data without requiring sequence alignment or even genome or locus assembly.

TopicContml is modular, and we have begun work to replace Contml with a network-generating package that may improve the analyses of such datasets by incorporating gene flow between species. Currently, for many datasets that do not suffer widespread introgression, TopicContml allows the analysis of many loci from many individuals that can be grouped, for example, into locations or species. We believe that TopicContml will become a valuable addition 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

Acknowledgments

We sincerely thank the reviewers for their constructive comments and suggestions, which significantly improved this manuscript. We acknowledge the contributions of our collaborators and are grateful to Subir Shakya and Tim Sackton for providing pre-publication access to PacBio reads from the two tinamou species. Additionally, we thank Liang Liu for assistance with interpreting the mammal dataset. We also thank the Louisiana State University Museum of Science for access to the *Tinamus guttatus* specimen used to generate PacBio sequence data, and the staff of the Museum of Comparative Zoology Ornithology Department for access to the jay specimens. Some simulations were conducted on the Research Computing Cluster at Florida State University, Tallahassee, and some data preprocessing was performed on the FASRC Cannon cluster, supported by the FAS Division of Science Research Computing Group at Harvard University.

Supplementary Material

Supplementary data is available at *SYSBIO* online.

Funding

This research was partly supported by the National Science Foundation grant DBI2019989 to P.B. and the National Institutes of Health grant 1R01HG011485 to S.V.E.

Data availability

- **Simulated Data**: Instructions for generating the simulated data are available at https://github.com/pbeerli/simulations_for_topiccontml.
- **Australian Treecreeper Dataset**: Available through Dryad (Edwards et al., 2022).
- **Mammal Dataset**: The aligned mammal dataset can be accessed at https://figshare.com/articles/cds_5162_zip/6031190 (Wu et al., 2018).
- **PacBio Bird Sequences**: The sources of PacBio long-read sequences from birds are detailed in Supplementary Section 4: PacBio Dataset, under Processing of PacBio Reads, and summarized in Supplementary Table S2. Datasets for the two tinamou species are available on Dryad at https://doi.org/10.5061/dryad.73n5tb36r.

.

References

Balaban M., Bristy N.A., Faisal A., Bayzid M.S., Mirarab S. 2022. Genome-wide alignment-free phylogenetic distance estimation under a no strand-bias model. *Bioinform. Adv.* 2(1):vbac055.

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

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

Cartwright R.A. 2005. Dna assembly with gaps (dawg): simulating sequence evolution. *Bioinformatics* 21(Suppl 3):iii31–8.

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

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

Chapus C., Dufraigne C., Edwards S., Giron A., Fertil B., Deschavanne P. 2005. Exploration of phylogenetic data using a global sequence analysis method. *BMC Evol. Biol.*, 5(1):63.

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

Cracraft J. 1986. Origin and evolution of continental biotas: Speciation and historical congruence within the australian avifauna. *Evolution*, 40(5):977–996.

Cracraft J. 1988. The major clades of birds. In: Benton M.J., editor. *The phylogeny and classification of the tetrapods, Volume 1: amphibians, reptiles, birds*, volume 35A of *Systematics Association Special*. Oxford: Clarendon Press. p. 339–361.

de Queiroz A., Gatesy J. 2007. The supermatrix approach to systematics. *Trends Ecol. Evol.* 22(1):34–41.

Deschavanne P.J., Giron A., Vilain J., Fagot G., Fertil B. 1999. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol. Biol. Evol.* 16(10): 1391–1399.

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

Du Y., Wu S., Edwards S.V., Liu L. 2019. The effect of alignment uncertainty, substitution models and priors in building and dating the mammal tree of life. *BMC Evol. Biol.* 19(1):203.

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

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

Edwards S.V., Tonini J.a.F.R., Mcinerney N., Welch C., Beerli P. 2022. Multilocus phylogeography, population genetics and niche evolution of Australian brown and black-tailed treecreepers (Aves: Climacteris) [Dataset]. https://doi.org/10.5061/dryad.bcc2fqzgt.

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

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

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

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

Felsenstein J. 1981. Evolutionary trees from gene frequencies and quantitative characters: finding maximum likelihood estimates. *Evolution*: 35:1229–1242.

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

Fofanov Y., Luo Y., Katili C., Wang J., Belosludtsev Y., Powdrill T., Belapurkar C., Fofanov V., Li T.-B., Chumakov S., Pettitt B.M. 2004. How independent are the appearances of n-mers in different genomes? *Bioinformatics* 20(15):2421–2428.

Foley N.M., Mason V.C., Harris A.J., Bredemeyer K.R., Damas J., Lewin H.A., Eizirik E., Gatesy J., Karlsson E.K., Lindblad-Toh K., Consortium Z., Springer M.S., Murphy WJ. 2023. A genomic timescale for placental mammal evolution. *Science* 380(6643): eabl8189.

Gatesy J., Baker R.H. 2005. Hidden likelihood support in genomic data: can forty-five wrongs make a right? *Syst. Biol.* 54(3):483–92.

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

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

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

Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y., Faircloth B.C., Nabholz B., Howard J.T., Suh A., Weber C.C., da Fonseca R.R., Li J., Zhang F., Li H., Zhou L., Narula N., Liu L., Ganapathy G., Boussau B., Bayzid M.S., Zavidovych V., Subramanian S., Gabaldón T., Capella-Gutiérrez S., Huerta-Cepas J., Rekepalli B., Munch K., Schierup M., Lindow B., Warren W.C., Ray D., Green R.E., Bruford M.W., Zhan X., Dixon A., Li S., Li N., Huang Y., Derryberry E.P., Bertelsen M.F., Sheldon F.H., Brumfield R.T., Mello C.V., Lovell P.V., Wirthlin M., Schneider M.P.C., Prosdocimi F., Samaniego J.A., Velazquez A.M.V., Alfaro-Núñez A., Campos P.F., Petersen B., Sicheritz-Ponten T., Pas A., Bailey T., Scofield P., Bunce M., Lambert D.M., Zhou Q., Perelman P., Driskell A.C., Shapiro B., Xiong Z., Zeng Y., Liu S., Li Z., Liu B., Wu K., Xiao J., Yinqi X., Zheng Q., Zhang Y., Yang H., Wang J., Smeds L., Rheindt F.E., Braun M., Fjeldsa J., Orlando L., Barker F.K., Jønsson K.A., Johnson

W., Koepfli K.-P., O'Brien S., Haussler D., Ryder O.A., Rahbek C., Willerslev E., Graves G.R., Glenn T.C., McCormack J., Burt D., Ellegren H., Alström P., Edwards S.V., Stamatakis A., Mindell D.P., Cracraft J., Braun E.L., Warnow T., Jun W., Gilbert M.T.P., Zhang G. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* 346(6215):1320–1331.

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

Li H.S. 2013. A fast and lightweight tool for processing fasta or fastq sequences. github. URL https://github.com/lh3/seqtk.

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

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

Liu L., Zhang J., Rheindt F.E., Lei F., Qu Y., Wang Y., Zhang Y., Sullivan C., Nie W., Wang J., Yang F., Chen J., Edwards S.V., Meng J., Wu S. 2017. Genomic evidence reveals a radiation of placental mammals uninterrupted by the kpg boundary. *Proc. Natl. Acad. Sci. U.S.A.* 114(35):E7282–E7290.

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

Minka T., Lafferty J. 2002. Expectation-propagation for the generative aspect model. In: *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc. p. 352–359.

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

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

Oliveros C.H., Field D.J., Ksepka D.T., Barker F.K., Aleixo A., Andersen M.J., Alström P., Benz B.W., Braun E.L., Braun M.J., Bravo G.A., Brumfield R.T., Chesser R.T., Claramunt S., Cracraft J., Cuervo A.M., Derryberry E.P., Glenn T.C., Harvey M.G., Hosner P.A., Joseph L., Kimball R.T., Mack A.L., Miskelly C.M., Peterson A.T., Robbins M.B., Sheldon F.H., Silveira L.F., Smith B.T., White N.D., Moyle R.G., Faircloth B.C. 2019. Earth history and the passerine superradiation. *Proc. Natl. Acad. Sci.* 116(16):7916–7925.

Ondov B.D., Treangen T.J., Melsted P., Mallonee A.B., Bergman N.H., Koren S., Phillippy A.M. 2016. Mash: fast genome and metagenome distance estimation using minhash. *Genome Biol.* 17:1–14.

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

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

Revell L.J. 2024. phytools 2.0: an updated R ecosystem forphylogenetic comparative methods (and other things). *PeerJ.* 12: e16505.

Röder M., Both A., Hinneburg A. 2015. Exploring the space of topic coherence measures. In: *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. WSDM '15 (ACM Press). p. 399–408.

Shedlock A.M., Botka C.W., Zhao S., Shetty J., Zhang T., Liu J.S., Deschavanne P.J., Edwards S.V. 2007. Phylogenomics of nonavian reptiles and the structure of the ancestral amniote genome. *Proc. Natl. Acad. Sci.* 104(8):2767–2772.

Shen W., Le S., Li Y., Hu F. 2016. Seqkit: a cross-platform and ultrafast toolkit for fasta/q file manipulation. *PLoS One* 11(10):e0163962.

Sievert C., Shirley K. 2014. LDAvis: a method for visualizing and interpreting topics. In: *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. Stroudsburg (PA): Association for Computational Linguistics. p. 63–70.

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

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

Van Etten J., Stephens T.G., Bhattacharya D. 2023. A k-mer-based approach for phylogenetic classification of taxa in environmental genomic data. *Syst. Biol.*, 72(5):1101–1118.

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

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

Wu S., Edwards S., Liu L. 2018. Genome-scale DNA sequence data and the evolutionary history of placental mammals. *Data Brief* 18: 1972–1975.

Wu S., Rheindt F.E., Zhang J., Wang J., Zhang L., Quan C., Li Z., Wang M., Wu F., Qu Y., Edwards S.V., Zhou Z., Liu L. 2024. Genomes, fossils, and the concurrent rise of modern birds and flowering plants in the late cretaceous. *Proc. Natl. Acad. Sci.* 121(8): e2319696121.

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

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