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, 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 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. We evaluated the performance of our method with biological and simulated data sets: a data set with 14 DNA sequence loci from 78-92 haplotypes from two Australian bird species distributed in 9 populations; a second data set of 5162 loci from 80 mammal species; and a third data set 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 sequence data and for k-mer data.

Key words: topic modeling; NLP; LDA; ContML; multilocus phylogeny; 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; Shedlock et al. 2007; Chapus et al. 2005; Deschavanne et al. 1999) 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). Alignment-free approaches for phylogeny inference have attracted considerable attention in recent years (Dylus et al. 2023; Aledo 2022; Bernard et al. 2021; Zielezinski et al. 2017). For example, several studies have shown that the alignment artifacts can significantly impact tree topology (Ogden and Rosenberg 2006; Wong et al. 2008; Du et al. 2019). Moreover, while accurate, alignment-based methods encounter various limitations, such as obstacles in whole genome comparisons due to 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. 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 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 introduced by Blei et al. (2003) within the context of unsupervised machine learning; it assumes that the documents consist of latent topics, each represented by the distribution of words. The goal is to uncover these topics by analyzing the words in the documents and essentially 'learning' the structure of the data. 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 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 derived from individuals grouped into populations or species.

Genome-wide data sets (sequences of whole genomes or multiple genes per species) are becoming increasingly prominent in inferring the evolutionary history of closely related species. Major approaches to analyzing multilocus data sets 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 (Gatesy and Baker 2005), (2) supertree methods to combine gene trees (de Queiroz and Gatesy 2007); (3) 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). We use probabilistic topic modeling to construct a species tree from multi-locus data 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 estimate 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 data set 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 (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 results of bootstrapping with SVDquartets (Chifman and Kubatko 2014) implemented in Paup* (Swofford 2003) (SVDquartets +Paup*) which is computationally efficient and able to perform well under a range of scenarios (Chou et al. 2015).

Description

TopicContml *Software*

TopicContml is a Python package 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 the LDA model for each locus (Fig. 1 left). For 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. (2) It uses these topic frequencies from multiple loci to estimate a species tree with CONTML (Fig. 1 right).

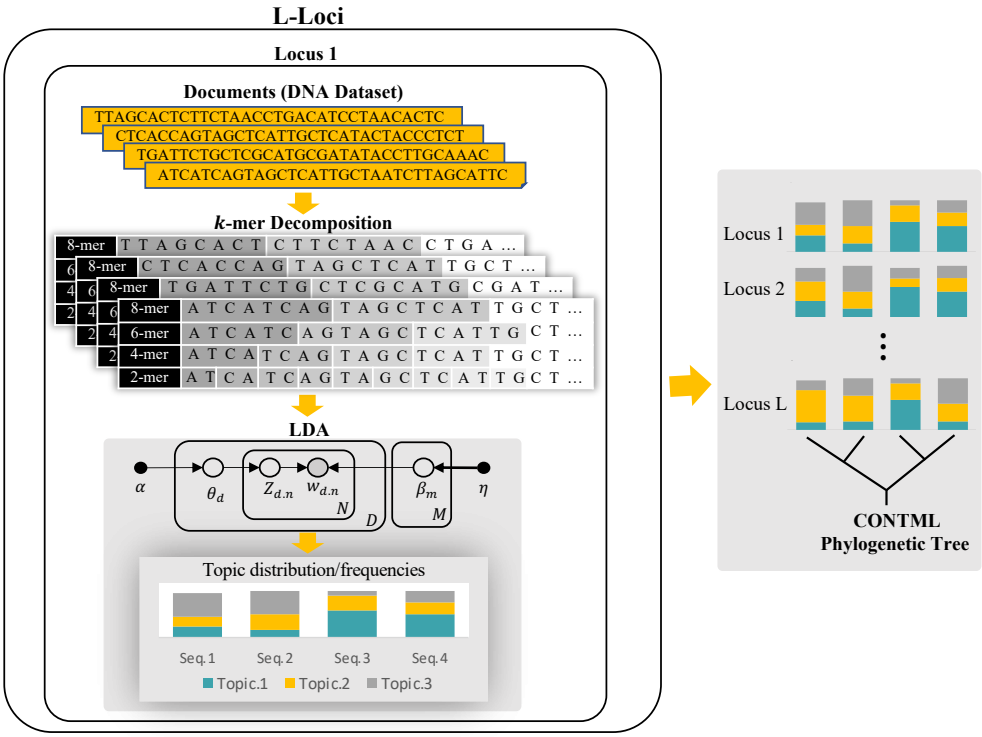

FIGURE 1. TOPICCONTML workflow to generate topic frequencies and the corresponding phylogeny.

K-mer decomposition.— In the context of NLP, tokenization involves dividing a text sequence into smaller units, referred to as tokens. Likewise, in bioinformatics, k-mers are k-length substrings of DNA or amino acid sequences, serving as tokens. We decompose the DNA sequence into non-overlapping k-mers, as shown in Fig. 1. The current program uses a set of four different values for k of 2, 4, 6, and 8, an approach that results in 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 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 *(Latent Dirichlet Allocation).*— The DNA sequences are assumed to be the corpus with D documents (sequences), 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), and then use these frequencies to generate a tree, as shown in Fig. 1.

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 that character state changes (nucleotide substitutions) 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: simulated data and real biological data. The simulated data were used to explore the effects of the number of loci and the divergence times within the true genealogy of 40 individuals from 4 species or populations. We used three biological data sets with very different features: (1) A 14-locus data set from nine closely related bird populations to discuss the accuracy of estimation, the effect of alignment, bootstrapping, and comparison to SVDquartets; (2) a vertebrate data set focusing on mammals with 90 species and 5162 loci to discuss missing data and the effect of aligned vs. non-aligned data. (3) a 7-species data set of 67316 autosomal loci and 4173 X-chromosome loci from mosquitoes to discuss the speed of operation;

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 the 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 (see Supplement Fig. S1). 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.

Real data.— Bird data set: The first data set of nine populations of two bird species, Australian brown treecreepers (*Climacteris picumnus*) and black-tailed treecreepers (*Climacteris melanurus*) (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 aligned sequences per species varied from 78 to

92 haplotypes. To obtain unaligned sequences, we removed all gaps in each sequence. We applied LDA to each locus across all 9 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 data set: The third data set contains 90 vertebrate species focusing on mammals with 5162 loci; the aligned data are available at https://figshare.com/articles/cds_5162_zip/6031190 (Wu et al. 2018) and were analyzed by Liu et al. (2017) using maximum likelihood. We ran this data with TopicConTML with several options: a. using aligned sequences, replacing the unspecified nucleotides (labeled 'N') with gaps. b. removing all gaps and all 'N' sites; c. removing all site columns in which any species had a gap or an 'N'. We then used the Robinson-Foulds distance to compare each phylogram with the maximum-likelihood tree and random trees. We also explored the run-times for this data set.

Mosquito data set: The second data set 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. https://doi.org/10.5061/dryad.tn47c). The data set contained 67317 autosomal loci and 4173 X-chromosome loci, each locus with different sequence lengths; for autosomes: average of 2299 base pairs with a range of 101 to 44205; and for X: average of 1167 base pairs with a range of 100 to 8010. Given the large number of loci, we were interested in the run-times for aligned vs. unaligned data.

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 the results of SVDquartets +Paup* analyzing the same data. This software infers species trees under the multispecies coalescent model using an invariants approach. 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 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.

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

Table 1. Simulated 4-population trees: TopicContml tree congruence with the true topology: Values are percentage of agreement; 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. S1)

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%.

Analysis of Real Data: Two Species with Multiple Populations

For the 'Bird' data set, 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 relationships within the species complex well; by contrast, lower numbers of topics do not yield a 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

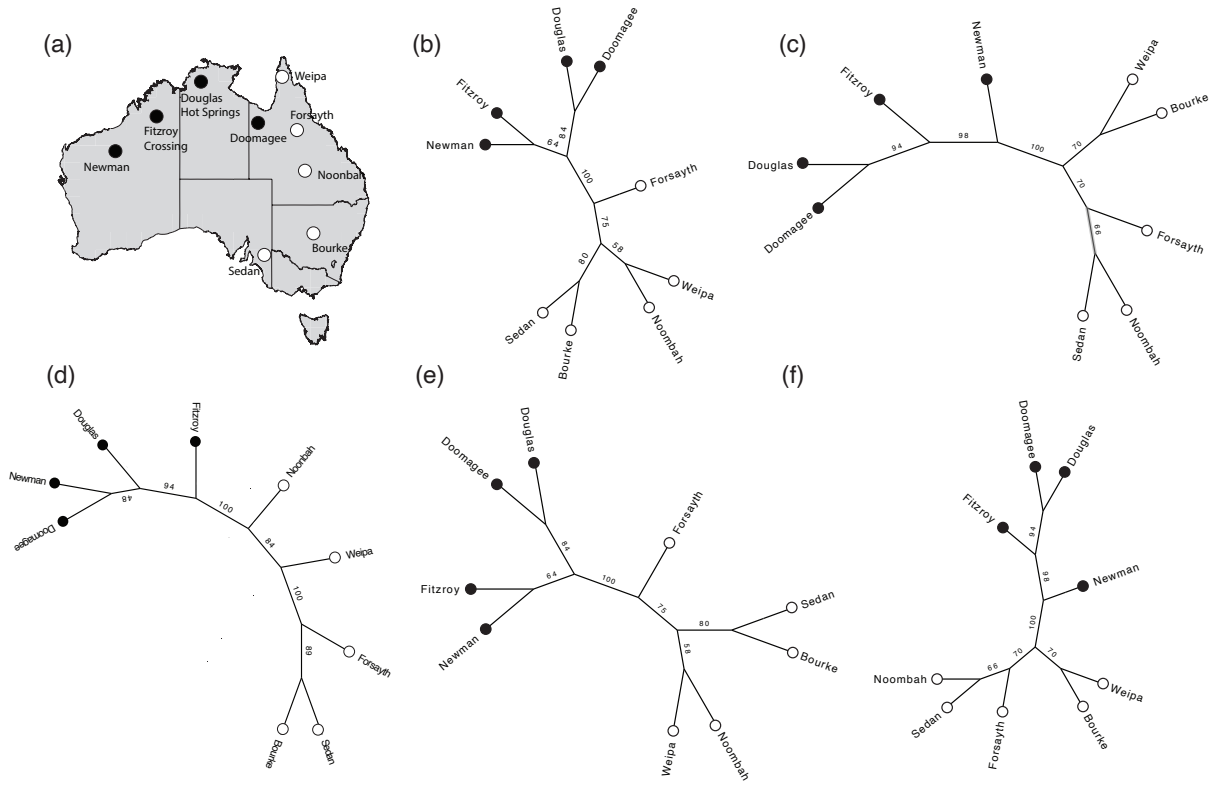

FIGURE 2. Relationship tree of 9 populations of two Australian treecreeper species reconstructed. (a) The map of Australia shows the locations; the black disks mark *Climacteris melanurus*, and the white disks mark *Climacteris picumnus* (Edwards et al. 2023). (b) The best phylogeny constructed by TopicContml from unaligned sequences with bootstrap support. (c) The best phylogeny constructed by TopicContml from aligned sequences. Bootstrap support comparison: (d) majority-consensus tree of aligned data analyzed by SVDquartets +Paup*. (e) majority-rule consensus tree of unaligned data by TopicContml. (f) majority-rule consensus tree of aligned data by TopicContml. For each bootstrap analysis, 1000 replicates were used. Values in the graphs are % support.

haplotypes as distinct taxa took several minutes (results not shown). Fig. 2b and Fig. 2c show the trees generated by TopicContml for the 9 locations where the birds were collected for unaligned and aligned sequences, respectively. Inspection of the clades reveals that our trees recover the expected geographic relationship within each species (Edwards et al. 2023; Cracraft 1986), and the locations are separated by species. Aligned and non-aligned trees are similar, although some locations, such as Sedan, Bourke, and Noonbah, are placed differently in the two trees.

Bootstrapping.— To compare the performance of TopicContml with SVDquartets, we applied both methods to the aligned 'Bird' data set and reported 50%

majority-rule consensus trees. For comparison, we also present the un-aligned bootstrap majority consensus tree. For each bootstrapped data set, we followed the same pipeline described in the TopicContml Software section, generating topic frequencies and corresponding phylogenies. Fig. 2e,f show majority-rule consensus trees with bootstrap support (1000 replicates). Fig. 2b,c show the bootstrap support superimposed onto the best trees from unaligned and aligned data. We compare the bootstrap results of TopicContml with those of SVDquartets +Paup* (Fig. 2d). There are conspicuous differences between SVDquartets and TopicContml results. Comparing them to the mapping of the locations on Fig. 2a, 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 two species. 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.

Analysis of Real Data: Multilocus Species Tree from Genome-wide Data

Mammal data set.— The topic modeling analyses of the mammal data set revealed several important issues of our approach. The data contained nucleotide characters from the set 'ACGT-N', treating the 'N' as a nucleotide led to the clustering of species that had many 'N' sites in common, which is undesirable; we replaced the 'N' with '-' which certainly distorts the original meaning of the gap sign, but reduced the vocabulary and improved the clustering of species that should be more closely related. For example, the marsupial *Macropus eugenii* (Tammar Wallaby) was located within the eutherian mammal clade before excising the 'N' sites (results not shown). Removing 'N' and assuming that they are equivalent to gaps led to an improved tree distance to the

maximum likelihood tree. Liu et al. (2017) used a maximum likelihood method on 4388 loci of 90 vertebrate species, and our comparisons are made against their maximum likelihood tree. Aligned sequences without 'N' were 96 steps from the Maximum likelihood tree using the Robinson-Foulds (RF) distance. Unaligned sequences without 'N' sites had a distance of 112. After removing all columns with gaps or 'N', our approach revealed a distance of 56 to the maximum likelihood tree. The RF distances of 1000 random trees to 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 considerably closer to the maximum likelihood tree than random trees, corroborating that our topic modeling approach reveals a phylogenetic signal. A tanglegram (Fig. 3) (Revell 2024) corroborates that our tree and the maximum likelihood tree deliver similar answers.

The runtime for 90 species and 5162 loci is substantial. The topic modeling only takes a few minutes to half an hour, but the ConTML run needs several hours for the 90 species.

Mosquito data set.— Recently, genome-wide data have led to the generation of well-resolved trees across the Tree of Life. While various methods are available, all require considerable time to generate a tree. Fontaine et al. (2015), along with a reanalysis of the mosquito data by Wen et al. (2016), show that a tree may not be the best way to visualize the relationships among the species. Since our method can only generate trees, we were interested in whether our method can recover the trees discussed in Fontaine et al. (2015) in a reasonable time. We evaluated 67517 autosomal loci of their *Anopheles gambiae* complex data set; our analysis derived the same topology (Supplement Fig. S4(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. S4(b)). Our runs of the aligned autosome and X-chromosome data took 22 minutes and 34 seconds, respectively. Both runs spend most of the time generating the topic

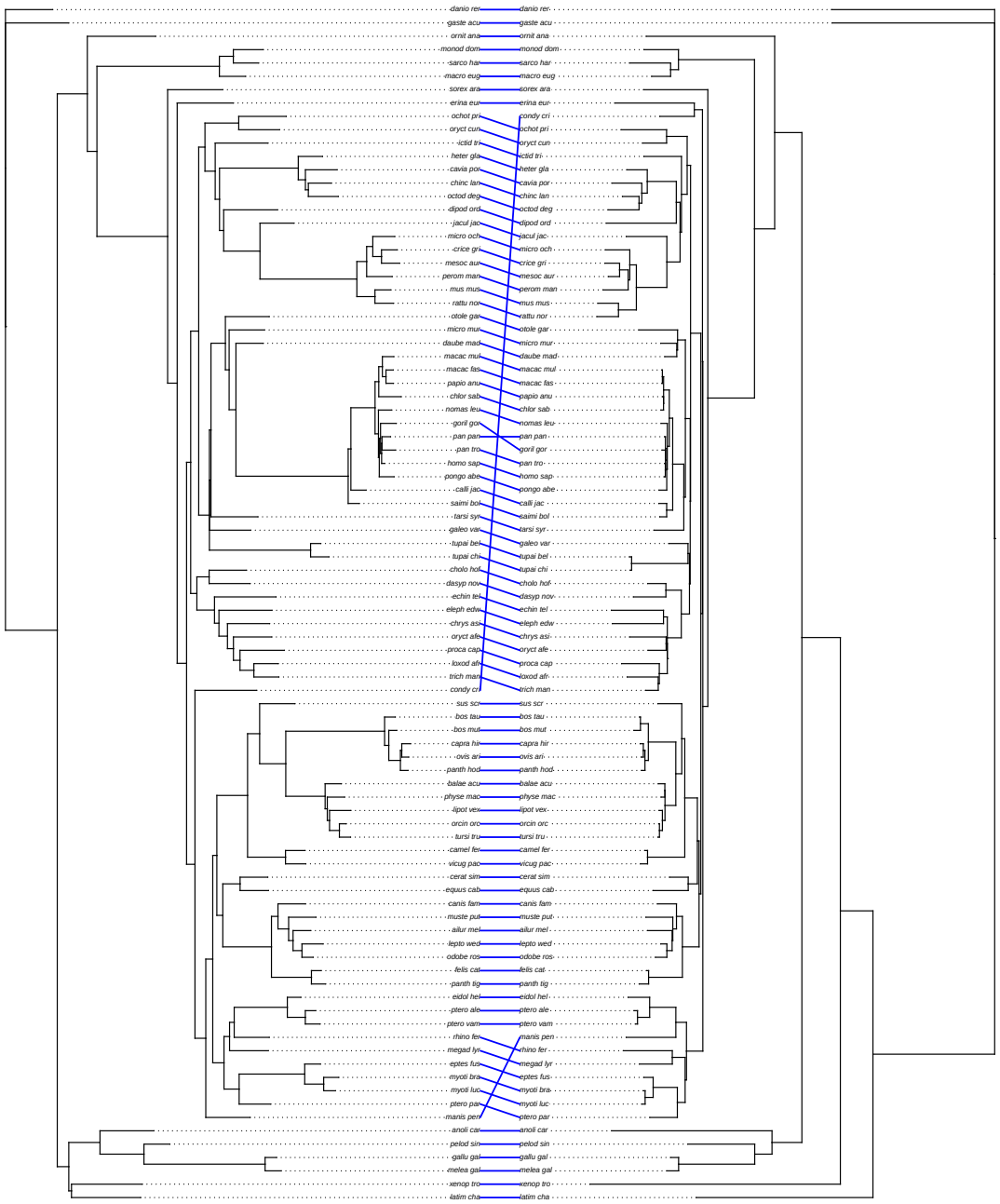

FIGURE 3. Tanglegram of the TopicContml tree Fig. S2 (left side) compared to the Maximum likelihood tree of Liu et al. (2017) (right side). The trees contain 90 vertebrate species. The left tree used TopicContml after all site columns containing gaps or 'N' were removed. The alphabetical list of the species names in the tree is in Supplement table 2.

frequencies (roughly about 1 to 2 seconds per 100 loci). In contrast, the unaligned data took roughly about 10× longer because the unaligned data led 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 conducted 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 data sets with a small number of species can be analyzed in a very short time on off-the-shelf computing devices.

Discussion

This study integrates 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 potentially 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 an arbitrary number of 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 to minutes on modern computer architectures and 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 LDA analyses delivered answers within seconds to minutes except for the mammal data set, where ContML took several hours to complete. The current version of ContML does not use parallelization, but this could be implemented and improve its runtime even for many species.

The simulated data experiments demonstrate that TopicContML running on

samples of many individuals from a few species can reveal the species topology without problems, even for moderate divergence time and small numbers of loci. When the divergence time approaches zero, our method finds the correct tree about 33% of all runs, delivering the statistically appropriate frequency of the correct topology. In contrast, the analysis of the real data sets was more complicated and we discuss the effect of structuring the sequence data into k-mers using the mammal data set. The Robinson-Foulds distances to the maximum likelihood tree (Liu et al. 2017) show clearly that analysis with more loci may not always lead to the best results. Our approach "`-gt rm_col`" removes all data columns where one or more species has a gap or an unknown nucleotide and has the smallest distance to the Liu tree. But in contrast to the fully aligned data set, it has discarded many loci because all their site columns with gaps and unspecified sites ('N') were deleted; the tree was then reconstructed with 763 loci instead of the available 5162 that include gaps for some individuals. When treating the data as unaligned, the resulting topology fits the least with the Liu tree. Many authors (cf. Edwards et al. 2002; Chan et al. 2014; Balaban et al. 2022; Van Etten et al. 2023; Aledo 2022; Dylus et al. 2023) declared that k-mers from unaligned data deliver good phylogenetic information, but our analysis shows that k-mers derived from unaligned sequences contain less information than sequences augmented with alignment information. We agree that unaligned data analysis is a great first and quick step to evaluate phylogenetic relationships, but if great accuracy is needed, then aligned sequences will be necessary. Still, given our reported speeds for our method, we believe that TopicContml is competitive with other methods.

Assessment of tree uncertainty is commonly done using bootstrapping the sequence data. In the context of unaligned sequences, this is difficult, and one needs to resort to bootstrapping the k-mers. 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 data sets 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 derived from these aligned sequences are more similar.

The very large number of autosomal loci for the mosquito data set also revealed that unaligned data increase the number of unique tokens for each individual leading to longer run times and also leading to longer branch lengths. Additionally, this data set 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 data sets.

For many data sets 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. It will become 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.

M. Balaban, N. A. Bristy, A. Faisal, M. S. Bayzid, and S. Mirarab. Genome-wide alignment-free phylogenetic distance estimation under a no strand-bias model. *Bioinform Adv*, 2(1):vbac055, 2022. ISSN 2635-0041 (Electronic); 2635-0041 (Linking). doi: 10.1093/bioadv/vbac055.

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.

G. Bernard, T. G. Stephens, R. A. González-Pech, and C. X. Chan. Inferring phylogenomic relationship of microbes using scalable alignment-free methods. In *Bacterial Pangenomics: Methods and Protocols*, pages 69–76. Springer, 2021.

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.

C. Chapus, C. Dufraigne, S. Edwards, A. Giron, B. Fertil, and P. Deschavanne. Exploration of phylogenetic data using a global sequence analysis method. *BMC Evolutionary Biology*, 5(1):63, 2005. doi: 10.1186/1471-2148-5-63. URL https://doi.org/10.1186/1471-2148-5-63.

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.

J. Cracraft. Origin and evolution of continental biotas: Speciation and historical congruence within the australian avifauna. *Evolution*, 40(5):977–996, 1986. ISSN 00143820, 15585646. URL http://www.jstor.org/stable/2408757.

A. de Queiroz and J. Gatesy. The supermatrix approach to systematics. *Trends in Ecology and Evolution*, 22(1):34–41, Jan 2007. ISSN 0169-5347 (Print); 0169-5347 (Linking). doi: 10.1016/j.tree.2006.10.002.

P. J. Deschavanne, A. Giron, J. Vilain, G. Fagot, and B. Fertil. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol Biol Evol*, 16(10):1391–1399, Oct 1999. ISSN 0737-4038 (Print); 0737-4038 (Linking). doi: 10.1093/oxfordjournals.molbev.a026048.

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.

Y. Du, S. Wu, S. V. Edwards, and L. Liu. The effect of alignment uncertainty, substitution models and priors in building and dating the mammal tree of life. *BMC Evolutionary Biology*, 19(1):203, 2019. doi: 10.1186/s12862-019-1534-9. URL https://doi.org/10.1186/s12862-019-1534-9.

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.

J. Gatesy and R. H. Baker. Hidden likelihood support in genomic data: can forty-five

wrongs make a right? *Systematic Biololgy*, 54(3):483–92, Jun 2005. doi: 10.1080/10635150590945368.

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.

L. Liu, J. Zhang, F. E. Rheindt, F. Lei, Y. Qu, Y. Wang, Y. Zhang, C. Sullivan, W. Nie, J. Wang, F. Yang, J. Chen, S. V. Edwards, J. Meng, and S. Wu. Genomic evidence reveals a radiation of placental mammals uninterrupted by the kpg boundary. *Proc Natl Acad Sci U S A*, 114(35):E7282–E7290, Aug 2017. ISSN 1091-6490 (Electronic); 0027-8424 (Print); 0027-8424 (Linking). doi: 10.1073/pnas.1616744114.

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.

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.

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.

L. J. Revell. phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things). *PeerJ*, 12:e16505, 2024. doi: 10.7717/peerj.16505.

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

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.

J. Van Etten, T. G. Stephens, and D. Bhattacharya. A k-mer-Based Approach for Phylogenetic Classification of Taxa in Environmental Genomic Data. *Systematic Biology*, 72(5):1101–1118, 06 2023. ISSN 1063-5157. doi: 10.1093/sysbio/syad037. URL https://doi.org/10.1093/sysbio/syad037.

S. Vinga and J. Almeida. Alignment-free sequence comparison 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.

S. Wu, S. Edwards, and L. Liu. Genome-scale dna sequence data and the evolutionary history of placental mammals. *Data Brief*, 18:1972–1975, Jun 2018. ISSN 2352-3409 (Print); 2352-3409 (Electronic); 2352-3409 (Linking). doi: 10.1016/j.dib.2018.04.094.

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, S. Vinga, J. Almeida, and W. M. Karlowski. Alignment-free sequence comparison: benefits, applications, and tools. *Genome Biol*, 18:1–17, 2017.

A. Zielezinski, H. Z. Girgis, G. Bernard, C.-A. Leimeister, K. Tang, T. Dencker, A. K. Lau, S. RÃ¶hling, 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

Simulated data

We used a coalescent tree simulator (*ms* (Hudson 2002)) and sequence data simulator (*migdata* (https://github.com/pbeerli/popsimulate)) to sequence data sets for 40 individuals from 4 populations

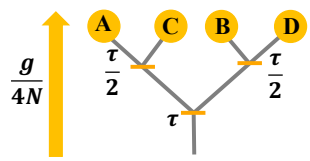

Figure S1. Population tree ((A,C),(B,D)) with split times. This tree was used to simulate trees with 10 individuals per population. The time scale is coalescent units, g is generations and N is the population size.

Mammal data analysis

We analyzed the mammal data with 5162 loci and present the results in Figure S2. We used a very restrictive scheme that removed all columns with one or more gaps and all columns with one or more unspecified nucleotide 'N'. This resulted in a tree generated using 763 loci because 4398 of the 5162 loci did not contain any data after the removal. This tree shows the result that is closest to the maximum likelihood tree presented by Liu et al. (2017) and has 56 Robinson-Foulds step difference. Most groups such as the bats, or whales and dolphins, are estimated well, but there are considerable differences in the placement of a few species (Figure 3; Figure S3 show the estimated trees using aligned data (5162 loci) and nonaligned data (5162 loci), both are not as consistent as the tree in Figure S2.

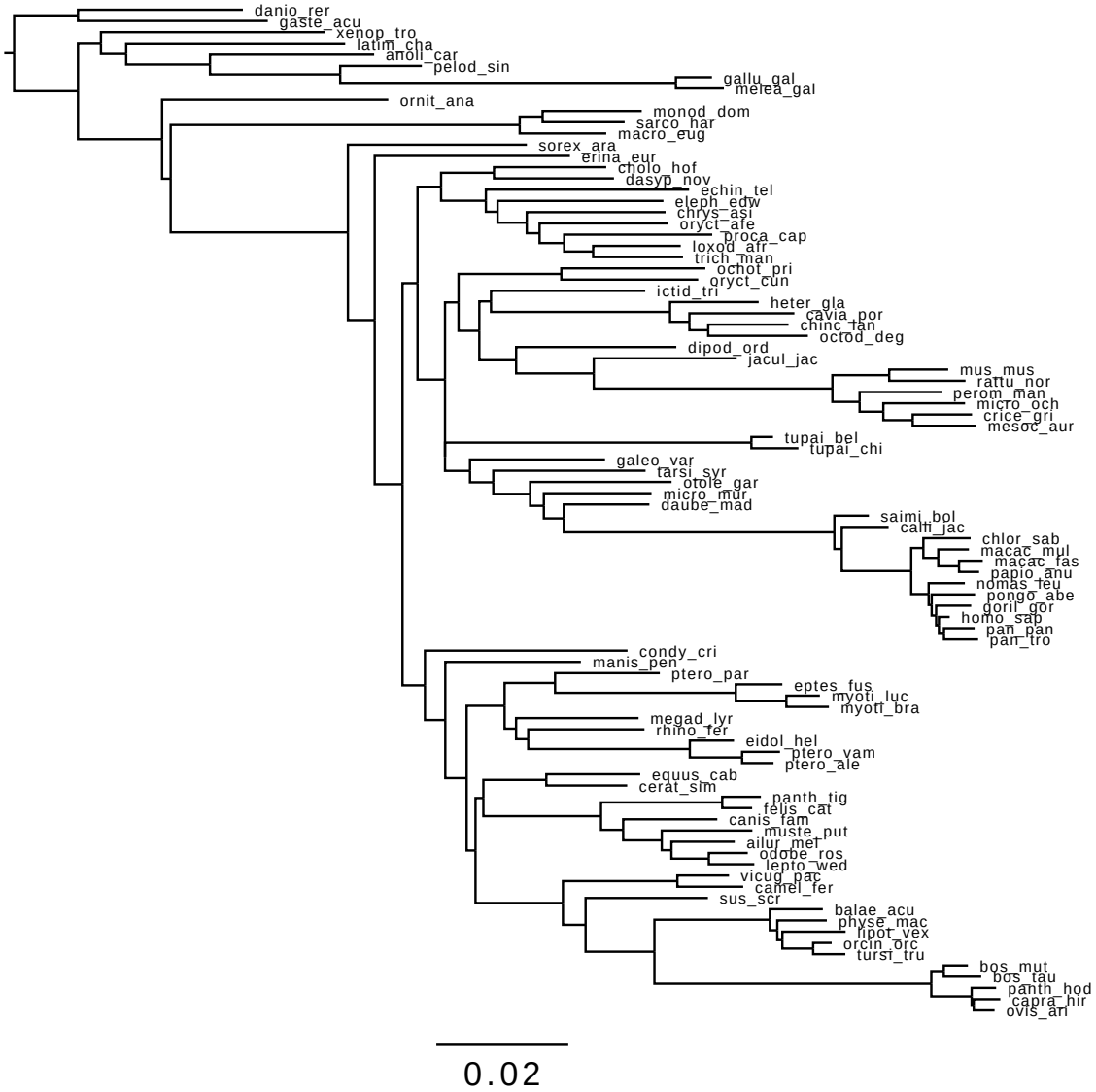

Figure S2. Phylogram of 90 vertebrate species generated with TopicContml using X loci after all site columns containing gaps or 'N' were removed. The alphabetical list of species names in the tree is in Table 2.

Table 2. Alphabetical list of species in figure S2

ailur_mel	*Ailurus fulgens* (Red Panda)
anoli_car	*Anolis carolinensis* (Carolina Anole)
balae_acu	*Balaenoptera acutorostrata* (Minke Whale)
bos_mut	*Bos mutus* (Wild Yak)
bos_tau	*Bos taurus* (Domestic Cattle)
calli_jac	*Callithrix jacchus* (Common Marmoset)
camel_fer	*Camelus ferus* (Wild Bactrian Camel)
canis_fam	*Canis familiaris* (Domestic Dog)
capra_hir	*Capra hircus* (Domestic Goat)
cavia_por	*Cavia porcellus* (Guinea Pig)
cerat_sim	*Ceratotherium simum* (White Rhinoceros)
chinc_lan	*Chinchilla lanigera* (Long-tailed Chinchilla)
chlor_sab	*Chlorocebus sabaeus* (Green Monkey)
cholo_hof	*Choloepus hoffmanni* (Hoffmann's Two-toed Sloth)
chrys_asi	*Chrysolophus amherstiae* (Lady Amherst's Pheasant)
condy_cri	*Condylura cristata* (Star-nosed Mole)
crice_gri	*Cricetulus griseus* (Chinese Hamster)
danio_rer	*Danio rerio* (Zebrafish)
dasyp_nov	*Dasypus novemcinctus* (Nine-banded Armadillo)
daube_mad	*Daubentonia madagascariensis* (Aye-aye)
dipod_ord	*Dipodomys ordii* (Ord's Kangaroo Rat)
echin_tel	*Echinops telfairi* (Small Madagascar Hedgehog)
eidol_hel	*Eidolon helvum* (African Straw-coloured Fruit Bat)
eleph_edw	*Elephas maximus* (Asian Elephant)
eptes_fus	*Eptesicus fuscus* (Big Brown Bat)
equus_cab	*Equus caballus* (Domestic Horse)
erina_eur	*Erinaceus europaeus* (European Hedgehog)
felis_cat	*Felis catus* (Domestic Cat)
galeo_var	*Galeopterus variegatus* (Flying Lemur)
gallu_gal	*Gallus gallus* (Red Junglefowl)
gaste_acu	*Gastereus aculeata* (Stickleback)
goril_gor	*Gorilla gorilla* (Western Gorilla)
heter_gla	*Heterocephalus glaber* (Naked Mole Rat)
homo_sap	*Homo sapiens* (Human)
ictid_tri	*Ictidomys tridecemlineatus* (Thirteen-lined Ground Squirrel)
jacul_jac	*Jaculus jaculus* (Lesser Egyptian Jerboa)
latim_cha	*Latimeria chalumnae* (Coelacanth)
lepto_wed	*Leptonycteris yerbabuenae* (Lesser Long-nosed Bat)
lipot_vex	*Lipotes vexillifer* (Yangtze River Dolphin)
loxod_afr	*Loxodonta africana* (African Elephant)
macac_fas	*Macaca fascicularis* (Crab-eating Macaque)
macac_mul	*Macaca mulatta* (Rhesus Macaque)
macro_eug	*Macropus eugenii* (Tammar Wallaby)
manis_pen	*Manis pentadactyla* (Chinese Pangolin)
megad_lyr	*Megaderma lyra* (Greater False Vampire Bat)
melea_gal	*Meleagris gallopavo* (Wild Turkey)
mesoc_aur	*Mesocricetus auratus* (Golden Hamster)
micro_mur	*Microtus muridae* (a species of vole)
micro_och	*Microcebus ochraceus* (Ochre Mouse Lemur)
monod_dom	*Monodelphis domestica* (Gray Short-tailed Opossum)
mus_mus	*Mus musculus* (House Mouse)
muste_put	*Mustela putorius* (European Polecat)
myoti_bra	*Myotis brandtii* (Brandt's Bat)
myoti_luc	*Myotis lucifugus* (Little Brown Bat)
nomas_leu	*Nomascus leucogenys* (Northern White-cheeked Gibbon)
ochot_pri	*Ochotona princeps* (American Pika)
octod_deg	*Octodon degus* (Degu)
odobe_ros	*Odocoileus rosenbergi* (White-tailed Deer)
orcin_orc	*Orcinus orca* (Killer Whale)
ornit_ana	*Ornithorhynchus anatinus* (Platypus)
oryct_afe	*Oryctolagus cuniculus* (European Rabbit)
oryct_cun	*Oryctolagus cuniculus* (European Rabbit)
otole_gar	*Otolemur garnettii* (Garnett's Galago)
ovis_ari	*Ovis aries* (Domestic Sheep)
pan_pan	*Pan paniscus* (Bonobo)
pan_tro	*Pan troglodytes* (Chimpanzee)
panth_hod	*Pantholops hodgsonii* (Tibetan Antelope)
panth_tig	*Panthera tigris* (Tiger)
papio_anu	*Papio anubis* (Olive Baboon)
pelod_sin	*Pelodiscus sinensis* (Chinese Softshell Turtle)
perom_man	*Peromyscus maniculatus* (Deer Mouse)
physe_mac	*Physeter macrocephalus* (Sperm whale)
pongo_abe	*Pongo abelii* (Sumatran Orangutan)
proca_cap	*Procavia capensis* (Rock Hyrax)
ptero_ale	*Pteropus alecto* (Black Flying Fox)
ptero_par	*Pteronotus parnellii* (Parnell's mustached bat)
ptero_vam	*Pteropus vampyrus* (Large Flying Fox)
rattu_nor	*Rattus norvegicus* (Norway Rat)
rhino_fer	*Rhinolophus ferrumequinum* (Greater horseshoe bat)
saimi_bol	*Saimiri boliviensis* (Bolivian Squirrel Monkey)
sarco_har	*Sarcophilus harrisii* (Tasmanian Devil)
sorex_ara	*Sorex araneus* (Eurasian Common Shrew)
sus_scr	*Sus scrofa* (Wild Boar)
tarsi_syr	*Tarsius syrichta* (Philippine Tarsier)
trich_man	*Trichechus manatus* (West Indian Manatee)
tupai_bel	*Tupaia belangeri* (Northern Treeshrew)
tupai_chi	*Tupaia chinensis* (Chinese Tree Shrew)
tursi_tru	*Tursiops truncatus* (Bottlenose Dolphin)
vicug_pac	*Vicugna pacos* (Alpaca)
xenop_tro	*Xenopus tropicalis* (Western Clawed Frog)

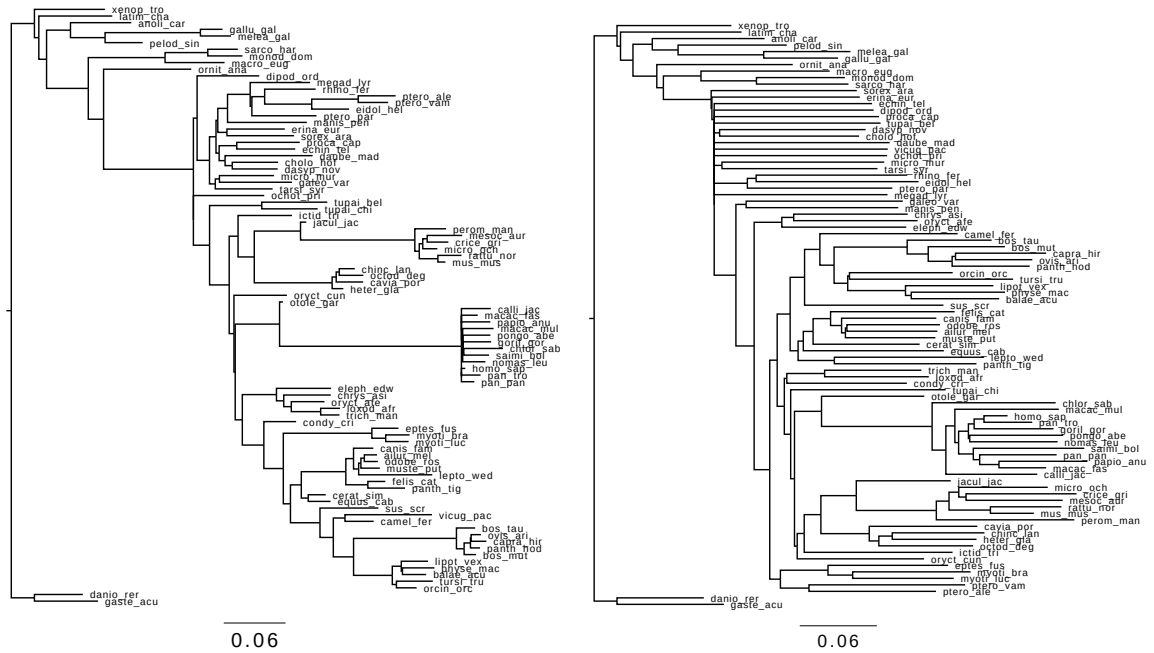

Figure S3. Phylograms of 90 vertebrate species generated with TopicContml using 5162 loci: left side: aligned data, 'N' sites are treated as gaps; right side: unaligned data, all gaps and 'N'sites were removed before the creation of the k-mers. The alphabetical list of species names in the tree is in Table 2.

Anopheles trees

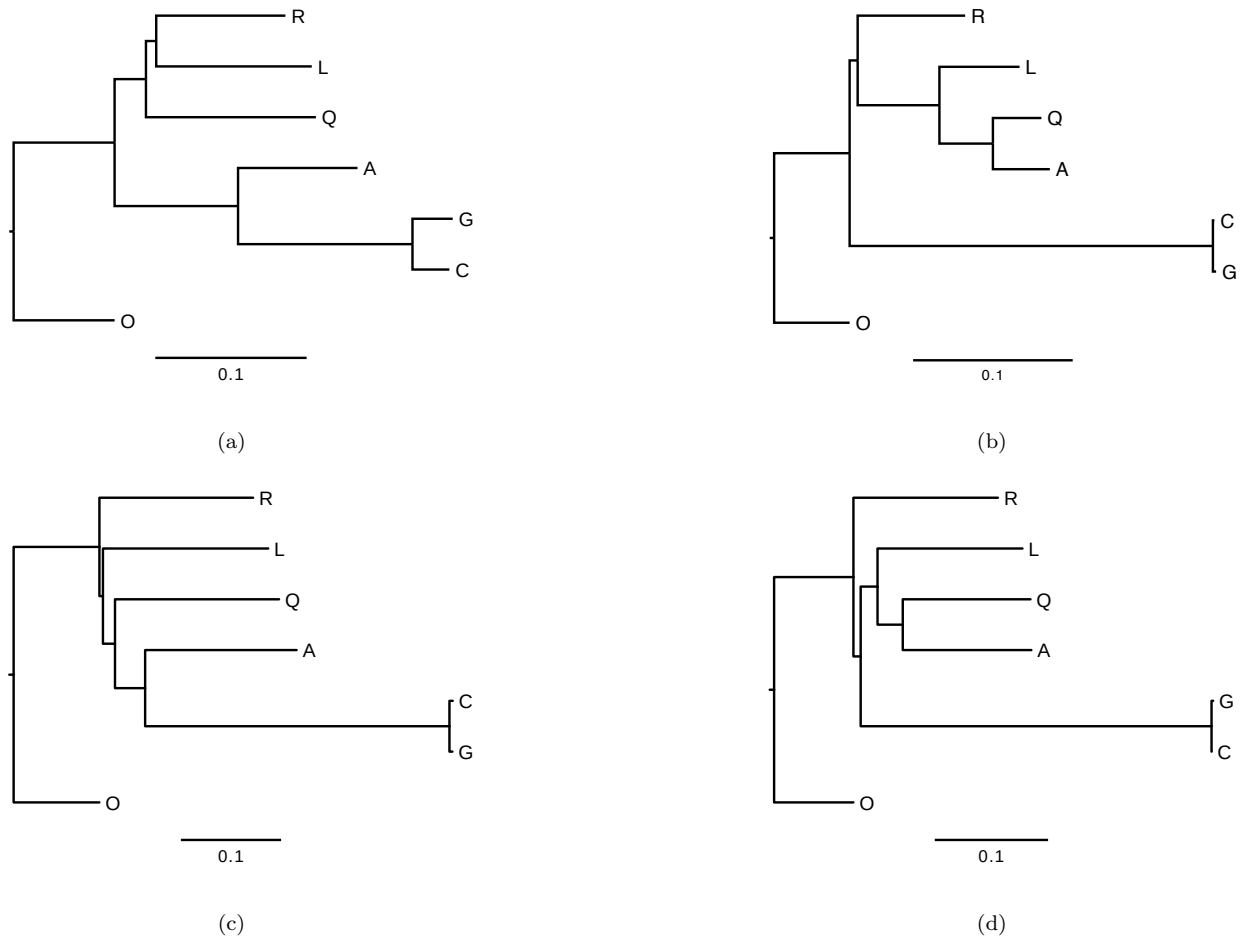

Figure S4. TopicContml analyses of 6 species in the *Anopheles gambiae* complex and an outgroup, using the data of Wen, D.; Yu, Y.; Hahn, M. W.; Nakhleh, L. [2016]. Data from: Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. Dryad. 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.