The Author(s) *BMC Genomics* 2016, **17**(Suppl 10):784
DOI 10.1186/s12864-016-3099-y

BMC Genomics

RESEARCH Open Access

Exploring phylogenetic hypotheses via Gibbs sampling on evolutionary networks

Yun Yu[1], Christopher Jermaine[1] and Luay Nakhleh[1,2*]

From 14th Annual Research in Computational Molecular Biology (RECOMB) Comparative Genomics Satellite Workshop
Montreal, Canada. 11-14 October 2016

Abstract

Background: Phylogenetic networks are leaf-labeled graphs used to model and display complex evolutionary relationships that do not fit a single tree. There are two classes of phylogenetic networks: Data-display networks and evolutionary networks. While data-display networks are very commonly used to explore data, they are not amenable to incorporating probabilistic models of gene and genome evolution. Evolutionary networks, on the other hand, can accommodate such probabilistic models, but they are not commonly used for exploration.

Results: In this work, we show how to turn evolutionary networks into a tool for statistical exploration of phylogenetic hypotheses via a novel application of Gibbs sampling. We demonstrate the utility of our work on two recently available genomic data sets, one from a group of mosquitos and the other from a group of modern birds. We demonstrate that our method allows the use of evolutionary networks not only for explicit modeling of reticulate evolutionary histories, but also for exploring conflicting treelike hypotheses. We further demonstrate the performance of the method on simulated data sets, where the true evolutionary histories are known.

Conclusion: We introduce an approach to explore phylogenetic hypotheses over evolutionary phylogenetic networks using Gibbs sampling. The hypotheses could involve reticulate and non-reticulate evolutionary processes simultaneously as we illustrate on mosquito and modern bird genomic data sets.

Background

Phylogenetic trees play a central role in evolutionary biology. A phylogenetic tree is most commonly inferred, directly or indirectly, from an alignment of sequences collected from a set of taxa of interest. The fundamental assumption underlying this inference step is that all characters in the alignment have evolved down a single tree in a strictly diverging manner. However, it is well established that different sites in the genome (and, different morphological characters) could evolve down different trees due to a host of biological processes (debate continues to rage regarding the size of genomic regions that could truly have a single underlying evolutionary tree [1, 2]). These processes can be divided into two categories: Treelike processes, which include incomplete lineage sorting (ILS) and gene duplication and loss (GDL), and reticulate, or non-treelike, processes, which include hybridization and horizontal gene transfer. From an evolutionary perspective, a major difference between these two categories is that the evolutionary history of the genomes is still adequately represented by a tree in the presence of treelike processes, whereas it is more appropriately represented by a network in the presence of reticulate processes. Since networks generalize trees, they can accommodate both categories of processes [3–5].

The term "phylogenetic network" encompasses many disparate models that allow topologies more general than trees. At the highest level of classification, phylogenetic networks can be grouped into data-display networks and evolutionary networks [6, 7]. A data-display network is a special type of undirected graphs that represents conflicts in the data, regardless of the causes of the conflict (the network could be treelike or reticulate) [7]. An evolutionary network is a special type of rooted, directed

*Correspondence: nakhleh@rice.edu
[1]Department of Computer Science, Rice University, 77005 Houston, Texas, USA
[2]Department of BioSciences, Rice University, 77005 Houston, Texas, USA

acyclic graphs that accommodates both treelike and reticulate evolutionary processes, yet distinguishes between the two in terms of the classification of its nodes [3]. Let us illustrate with an example of four sequences of two sites each, TT, TG, GG, and GT, from four taxa A, B, C, and D, respectively. Assuming, for example, that no recurrent or parallel mutation occurred at any of the two sites, then these four sequences cannot be modeled with a single tree since the two sites give conflicting signals (the first sites groups A and B together, while the second site groups B and C together). A data-display network of these four sequences is shown in Fig. 1a. If we cut the two horizontal lines in the box, we obtain a split that groups A and B together and groups C and D together. If we cut the two vertical lines in the box, we obtain a split that groups B and C together and groups A and D together. In this manner, a data-display network can represent a set of conflicting splits (and trees). However, it is important to emphasize that these networks are analyzed and interpreted in a special way: To obtain a split, or bipartition, of the data, only maximal sets of parallel lines (edges) can be cut.

An evolutionary network of the same four sequences is shown in Fig. 1b. This network gives an explicit model of the evolutionary history with a precise interpretation of the processes (in this illustration, it is a reticulation event, e.g., hybridization, that involves taxon B). Needless to say, the conflict in the data could be due to a recurrent mutation, e.g., at the second site, and the data could fit a tree (Fig. 1c). However, it is important to point out that these structures are used for modeling genome-wide incongruences, where processes such as ILS, GDL, etc., are at play in many data sets.

The efficiency with which data-display networks could be reconstructed and the availability of a popular tool, SplitsTree [8], that provides user-friendly implementation of several algorithms for their inference, makes them commonly used for exploring data. Evolutionary networks, on the other hand, have been used to incorporate statistical models such as the multispecies coalescent [9, 10] and, as a result, their statistical inference [11] is currently computationally prohibitive except for small data sets. Therefore, evolutionary networks have not been used for exploring data.

In this paper, we develop a framework for exploring evolutionary hypotheses, including treelike ones such as different tree rootings, via a novel application of Gibbs sampling to evolutionary networks. While in this work we focus on the multispecies coalescent, thus allowing to explore hypotheses that involve ILS and reticulations, our model could be extended to incorporate statistical models of other processes, such as GDL. We demonstrate the application of our framework to explore evolutionary hypotheses that arose in two recent studies of genomes of mosquitos [12] and modern birds [13]. Furthermore, we study the performance of our framework on simulated data to assess its accuracy. While exploration of evolutionary processes using this statistical framework is still more computationally expensive than data-display networks, it results in more specific hypotheses and allows for explicit incorporation of evolutionary models of genes and genomes. The method is implemented and publicly available as part of the PhyloNet software package [11].

Method

The posterior of phylogenetic networks and their parameters

A (binary) *phylogenetic network* [3] N on set of taxa $\mathscr{X}$ is a rooted, directed, acyclic graph whose leaves are bijectively labeled by $\mathscr{X}$ and whose every internal node v (except for the root which has $indeg(v) = 0$) has $indeg(v) = 1$ and $outdeg(v) = 2$ (*tree node*), or $indeg(v) = 2$ and $outdeg(v) = 1$ (*reticulation node*). Here, *indeg* and *outdeg* denote the in- and out-degree of a node, respectively. We denote by $E(N)$ the set of edges of N. The phylogenetic network N has branch lengths $\lambda : E(N) \rightarrow \mathbb{R}^+$, where $\lambda(e)$ is the length of edge e in coalescent units. Furthermore, associated with each reticulation edge e is a value $\gamma_e \in [0, 1]$, such that if two reticulation edges e_1 and e_2

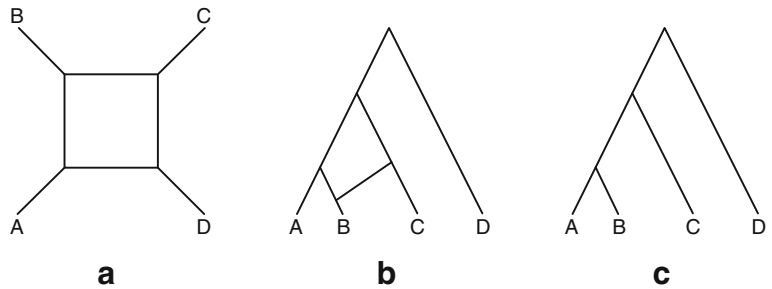

Fig. 1 Three different models of four sequences of two sites each, TT, TG, GG, and GT, from four taxa A, B, C, and D, respectively. **a** A data-display network that accommodates the two conflicting splits AB|CD and BC|AD. **b** An evolutionary network that explicitly models a reticulation event involving taxon B. **c** A tree model that would fit the data if, for example, a recurrent mutation occurred at the second site

are incident into the same reticulation node, then $\gamma_{e_1} + \gamma_{e_2} = 1$. These γ values represent the *inheritance probability* associated with a reticulation node. Throughout this paper, we denote by θ the parameters of a phylogenetic network N which include both the branch lengths λ and inheritance probabilities γ of N. If N has k_1 edges and k_2 reticulation nodes, then θ is of size $k_1 + k_2$. The network topology along with θ define a generative model of gene trees in the presence of reticulation under the multispecies network coalescent model [9, 10].

Given a set of gene trees $\mathcal{G}$ from a set of independent loci and a phylogenetic network N, the posterior distribution of N and θ is given by

$$p(N,\theta|\mathcal{G}) \propto p(\mathcal{G}|N,\theta)p(N,\theta) = p(N,\theta)\prod_{g\in\mathcal{G}} p(g|N,\theta) \tag{1}$$

where the product over the gene trees is based on the assumption that the loci are independent. The probability density function (PDF) $p(g|N,\theta)$ when the gene tree is given by its topology and branch lengths was derived in [10] and the probability mass function (PMF) for gene tree topologies alone was derived in [9] and an efficient algorithm for its computation was developed in [14]. As estimating gene tree branch lengths is challenging and negatively affects parameter estimation [15], we focus in this work on the scenario where the data consist of gene tree topologies alone. However, the method applies in a straightforward manner to data that consist of gene trees with branch lengths, with the only difference from what we describe below being the use of the PDF, rather than PMF, in computing the likelihood.

In this paper, we focus on (evolutionary) phylogenetic networks as an exploratory tool. That is, scenarios we envision are ones where the practitioner proposes a network topology and uses the gene tree data to explore the posterior of the network's parameters to determine which edges are supported by the data. Therefore, the distribution of interest in this case is the posterior on the parameters θ for a fixed network N. We illustrate the exploratory power of the method on two recently available biological data sets in the Results section. We now describe how to apply Gibbs sampling to obtain a posterior distribution of a given network's parameters.

A Gibbs sampling approach

Gibbs sampling [16] is a Markov chain Monte Carlo (MCMC) algorithm commonly used for sampling from the posterior distribution of a parameter set such as θ. The algorithm begins with an initialization $\theta^{(0)}$. Then, some subset of the parameters θ is updated by sampling from the target distribution of the subset conditioned on the known values of all other parameters. This is repeated for different subsets until convergence. In the particular version of Gibbs sampling we consider, the algorithm proceeds in a series of iterations, where in each iteration, each parameter θ_i is updated in sequence. That is, in each iteration, a value of parameter θ_i is sampled from the conditional distribution $p(\theta_i|\theta_{\setminus i},\mathcal{G},N)$, where $\theta_{\setminus i}$ denotes that the values of all parameters in θ are fixed except for θ_i. for simplicity. Note that when $\theta_{\setminus i}$ is fixed we have

$$\begin{aligned} p(\theta_i|\theta_{\setminus i},\mathcal{G},N) &= \frac{p(\theta,\mathcal{G},N)}{p(\theta_{\setminus i},\mathcal{G},N)} \\ &= \frac{p(\mathcal{G}|N,\theta)p(N,\theta)}{p(\theta_{\setminus i},\mathcal{G},N)} \propto p(\mathcal{G}|N,\theta)p(N,\theta). \end{aligned} \tag{2}$$

Thus, when sampling from $p(\theta_i|\theta_{\setminus i},\mathcal{G},N)$, we can calculate $p(\mathcal{G}|N,\theta)p(N,\theta)$ with only θ_i changing and sample from it. For the prior $p(N,\theta)$, since N is fixed, we focus on $p(\theta)$. For branch lengths, we use the exponential distribution with parameter $\lambda = 1$, which is a standard prior [17]. For the inheritance probabilities, we assume the U-shaped Beta distribution with parameters $\alpha = \beta = 0.1$ to reflect the belief that a majority of the reticulation edges do not exist in reality. For both the branch lengths and inheritance probabilities, any prior could be used without modifying the algorithm.

The major challenge in implementing the Gibbs sampler in our case is that it is very hard to sample from the conditional distribution. To overcome this challenge we use rejection sampling. We implement an algorithm that progressively builds a more accurate, step-wise over-approximation of the posterior for use in the rejection sampling. Suppose we are sampling θ_i whose range is $[x_1^l, x_1^h]$. The rejection sampling starts with a uniform envelope whose height is $y_1 = \max p(\mathcal{G}|N,\theta)p(N,\theta)$ computed using finite difference-based gradient descent (the dotted line in Fig. 2a) with θ_i set to $x_1 = \arg\max p(\mathcal{G}|N,\theta)p(N,\theta)$. If no sample is accepted within some preset number of trials (*maxFailure* in the algorithm below), the envelope is adjusted by breaking it down into more rectangles, such as in Fig. 2b, and the rejection sampling is repeated. If the sampling fails again, the envelope is further refined as in Fig. 2c. This process is repeated until one sample is successfully obtained.

Algorithm 1 gives the pseudo-code of one iteration of the Gibbs sampler. The input to each iteration is the set of gene trees $\mathcal{G}$, a phylogenetic network topology N, the values of the parameters θ from the previous iteration, the number of trials before adjusting the envelope *maxFailure*, the bounds within which to sample parameter values x_1^l and x_1^h, and thresholds τ, ϵ, and δ used in

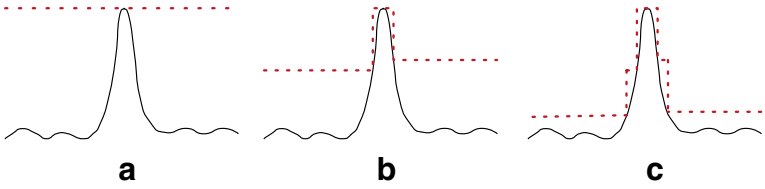

Fig. 2 Building envelopes for rejection sampling. The *black curves* are the distribution we want to sample from. The *red dotted lines* correspond to envelopes built for rejection sampling. **a** The initial envelope. **b** The adjusted envelope after the initial envelope fails to produce an accepted sample. **c** The envelope is further refined

the envelope construction. In our analyses here, we used $maxFailure = 10, \tau = 1/100, \epsilon = 0.001$, and $\delta = 0.2$. For the bounds on the parameter values we used the range $[0.001, 6]$ for branch lengths and the range $[10^{-6}, 1 - 10^{-6}]$ for inheritance probabilities.

Algorithm 1: GibbsSamplingIteration

Input: A set of gene trees $\mathcal{G}$, a phylogenetic network N, current values of parameter vector θ, $maxFailure \in \mathbb{N}$ and $\tau, \epsilon, \delta, x_1^l, x_1^h \in \mathbb{R}$
Output: A sample of values of the parameter vector θ
for $i = 1$ **to** $|\theta|$ **do**
  Let x_1^l and x_1^h be the lower and upper bounds of θ_i, respectively;
  $\theta_i \leftarrow x_1 \leftarrow \arg\max_{\theta_i} p(\mathcal{G}|N,\theta)p(N,\theta)$;
  $y_1 \leftarrow p(\mathcal{G}|N,\theta)p(N,\theta)$; `// The height of the envelope`
  $ns \leftarrow 1$; `// The number of steps in the envelope`
  $success \leftarrow false$; `// Indicates whether a sample has been accepted`
  while !$success$ **do**
    repeat
      Sample θ_i in the range $[x_l, x_r]$ where $x_l = \min_{j \le ns} x_j^l$ and $x_r = \max_{j \le ns} x_j^r$;
      Let k be the index such that $\theta_i \in [x_k^l, x_k^h]$;
      Sample $\rho \sim U(0,1)$;
      if $\rho < \frac{p(\mathcal{G}|N,\theta)p(N,\theta)}{y_k}$ **then**
        $success \leftarrow true$;
      end if
    until *maxFailure attempts have been made or success is true*;
    if !$success$ **then**
      if $x_1 - x_1^l > \epsilon$ **then** `// refine the left end of the envelope`
        for $j = ns$ **to** 1 **do**
          $x_{j+1}^l \leftarrow x_j^l, x_{j+1}^h \leftarrow x_j^h, x_{j+1} \leftarrow x_j, y_{j+1} \leftarrow y_j$;
        end for
        do
          $\delta \leftarrow \delta/2$;
          $x_1^h \leftarrow x_2 - \delta$;
          $y_1 \leftarrow p(\mathcal{G}|N,\theta)p(N,\theta)$ with $\theta_i \leftarrow x_1^h$;
        while $x_1^h < x_1^l$ *or* $\frac{y_1}{y_2} < \tau$;
        $x_1 \leftarrow x_1^h, x_2^l \leftarrow x_1^h, ns \leftarrow ns + 1$;
      end if
      if $x_{ns}^h - x_{ns} > \epsilon$ **then** `// refine the right end of the envelope`
        $x_{ns+1}^h \leftarrow x_{ns}^h, y_{ns+1}^h \leftarrow y_{ns}^h$;
        do
          $\delta \leftarrow \delta/2$;
          $x_{ns+1}^l \leftarrow x_{ns} + \delta$;
          $y_{ns+1} \leftarrow p(\mathcal{G}|N,\theta)p(N,\theta)$ with $\theta_i \leftarrow x_{ns+1}^l$;
        while $x_{ns+1}^h < x_{ns+1}^l$ *or* $\frac{y_{ns+1}}{y_{ns}} < \tau$;
        $x_{ns+1} \leftarrow x_{ns+1}^l, x_{ns}^h \leftarrow x_{ns+1}^l, ns \leftarrow ns + 1$;
      end if
    end if
  end while
end for

The Gibbs sampler performs each iteration described in Algorithm 1 a *maxIterations* number of times, and then collects samples every *sampleInterval* iterations after an initial burn-in period of *burnin* iterations. For all analyses we conducted below, we used $maxIterations = 11000$, $burnin = 1000$, and $sampleInterval = 100$.

Using pseudo-likelihood

The bottleneck of our method in terms of scalability results from computing the likelihood function $p(\mathcal{G}|N,\theta)$. In every iteration of the Gibbs sampling, the likelihood $p(\mathcal{G}|N,\theta)$ is computed repeatedly when building envelopes and conducting rejection sampling. This computation is very expensive, which makes the method infeasible for large data sets (such as the avian data set below). Pseudo-likelihood of phylogenetic networks was recently introduced [18] and its computation is very efficient as it is based on the probabilities of rooted triplets (rooted trees with three leaves) rather than full gene trees. The main issue with using the pseudo-likelihood is that it might result in indistinguishability of different parameter values, as discussed in [18].

Network inference

Wen et al. [19] recently introduced a Bayesian Markov chain Monte Carlo (MCMC) method for sampling the posterior of phylogenetic networks. Their work entails walking the space of phylogenetic network topologies, branch lengths and inheritance probabilities. One way to use the method presented here to infer, rather than explore, a phylogenetic network is by using an overly complex network that, desirably, contains within it the true network, and then apply our method to obtain a posterior distribution of its parameters. The major bottleneck in this case would be computing the PMF, as its computational complexity explodes as the number of reticulations increases. An advantage of the approach, though, would be avoiding the sampling, comparison, and summarization of the network topologies, all of which are very challenging as discussed in [19]. A disadvantage, though, is that evolutionary relationships not present in the network being analyzed will not be recovered or assessed in the analysis.

Results and discussions

Performance on simulated data

To study the accuracy of our sampler, we consider two simulated data sets. The phylogenetic topologies and associated parameters are modeled after the topologies and parameters of the mosquito data set of [12]. In the first data set, the model species phylogeny is a tree, shown in Fig. 3a. All branch lengths are set to 1 coalescent unit. We use our method to explore several phylogenetic hypotheses, represented in the network shown in Fig. 3b. Through this species network we can test two treelike issues that do not involve true reticulation. One is the rooting of the species tree. The reticulation on the top indicates two different rootings. One splits $\{C, G\}$ and $\{A, Q, L, R\}$ and the other splits $\{R\}$ and $\{A, Q, L, C, G\}$. The second issue we can test through this network is the location of Q: whether it should be grouped with A or R. It is captured by the lower reticulation.

To assess the performance of our method, we used ms [20] to simulate 50, 100, 200, 500, 1000 and 2000 gene trees within the branches of the true species phylogeny. For each number of gene trees, 30 data sets were generated. Then down each gene tree we simulated sequences of lengths 1000 under the general time-reversible (GTR) model using seq-gen [21]. The population mutation rate was set to 0.036. The base frequencies of the nucleotides A, C, G and T were set to 0.2112, 0.2888, 0.2896 and 0.2104, respectively. The relative rates of substitutions were set to 0.2173, 0.9798, 0.2575, 0.1038, 1 and 0.2070. Finally, gene trees were reconstructed using RAxML [22] and then rooted at the outgroup. RAxML was run five times for each sequence alignment to obtain the estimated gene tree.

We ran our method on the species network in Fig. 3b along with true gene trees and reconstructed gene trees. We used full-likelihood to compute $p(\mathcal{G}|N, \theta)$ in Eq. (1). After we collected samples from the Gibbs sampler, we pruned the collected networks by removing all reticulations with inheritance probabilities lower than 0.01. The results are shown in Fig. 3c. The posterior probabilities of true networks were calculated as the proportion of the true networks appearing in the final set of pruned networks. The red and blue boxes in the figure represent results from true gene trees and reconstructed gene trees, respectively. As the results demonstrate, as more gene trees are used in the input, the true phylogeny is more likely to be sampled. Furthermore, the results from reconstructed gene trees and results from true gene trees differ only slightly, demonstrating robustness to gene tree estimation errors.

In the second simulated data set, we tested the case where the model species phylogeny has reticulations. We conducted simulations on the true network with one reticulation, shown in Fig. 4a. All branch lengths are set to 1 coalescent unit, and the inheritance probability is set to 0.2. Our exploratory phylogenetic hypothesis is the species network shown in Fig. 4b), which contains two scenarios for testing. One is whether gene flow is from Q to R or R to Q, and the other is the location of A or whether there is gene flow from the ancestor of C and G to A. To test whether our method can recover the true gene flow, we used the same settings as in the first case to generate true gene trees and reconstructed gene trees and then ran our method on those gene trees. The results are shown in Fig. 4c. As the results show, the posterior probabilities of the true network increase with the number of gene trees

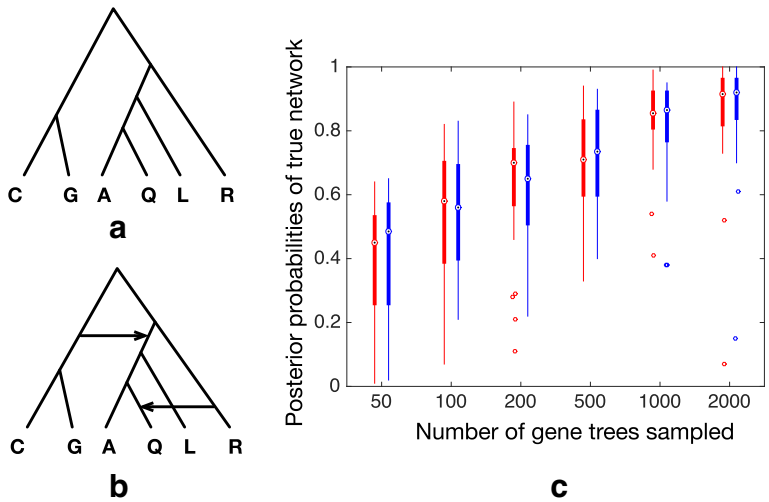

Fig. 3 Simulation study 1. **a** The true species phylogeny (a tree) with all internal branches set to 1 coalescent unit. **b** The species network fed into our method. **c** The results where *red boxes* represent results from using true gene trees and *blue boxes* represent results from using reconstructed gene trees

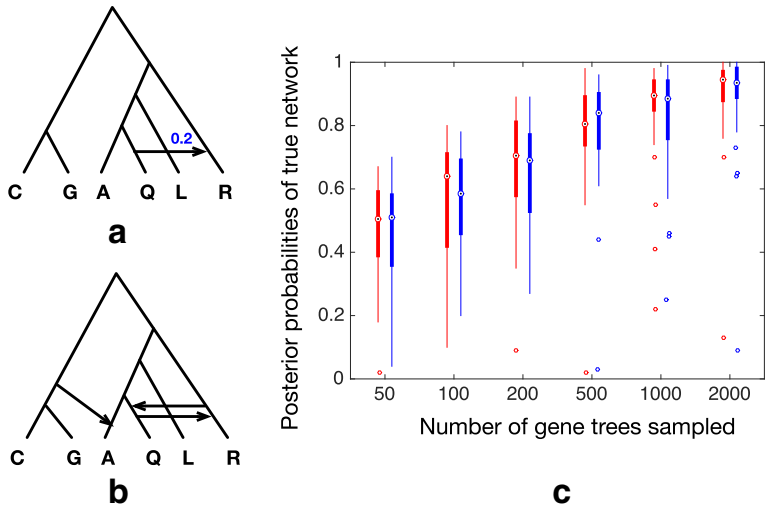

Fig. 4 Simulation study 2. **a** The true species phylogeny (a network with one reticulation) with all internal branches set to 1 coalescent unit and inheritance probability set to 0.2. **b** The species network fed into our method. **c** The results where *red boxes* represent results from using true gene trees and *blue boxes* represent results from using reconstructed gene trees

sampled. Also, the results from reconstructed gene trees only differ slightly when comparing to the results from true gene trees.

As we discussed above, in order to improve the scalability, we proposed to use pseudo-likelihood instead of full-likelihood to calculate $p(\mathcal{G}|N,\theta)$ in Eq. (1). We studied the performance of our method using pseudo-likelihood versus using full-likelihood when the number of gene trees is large. More specifically, for both simulation cases we studied (Figs. 3 and 4), we ran our method using pseudo-likelihood instead of full-likelihood on data sets of 2000 true gene trees. Results are shown in Fig. 5. We can see that the posterior probabilities of the true networks from using pseudo-likelihood are slightly lower than those from using full-likelihood when both of them use 0.01 as threshold to prune networks (remove reticulation edges whose inheritance probabilities are lower than 0.01). However, if we change the threshold slightly to 0.015, then the results from using pseudo-likelihood are almost the same as results from using full-likelihood.

Analysis of a mosquito data set

In a recent study, Fontaine et al. [12] conducted phylogenomic analysis of six members of the *Anopheles gambiae* species complex, including *An. gambiae* (*gam*), *An. coluzzii* (*col*), *An. arabiensis* (*ara*), *An. quadriannulatus* (*qua*), *An.merus* (*mer*) and *An. melas* (*mel*). The authors reported extensive incongruence among gene trees due to both incomplete lineage sorting and introgression and presented a reticulate evolutionary history of this group, which is the network shown in Fig. 6a with gene flow between *An. arabiensis* and the ancestor of *An. gambiae*

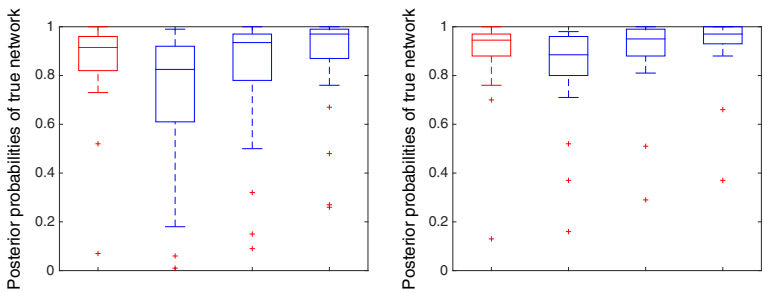

Fig. 5 Using pseudo-likelihood versus full-likelihood. *Left and right* panels correspond to results of case 1 (Fig. 3) and case 2 (Fig. 4), respectively. For both, 2000 true gene trees were used. In each panel, *red box* represents results from using full-likelihood when threshold 0.01 was used to prune networks. *Blue boxes* represent results from using pseudo-likelihood, while *three blue* boxes from left to right correspond to using threshold 0.01, 0.015 and 0.02, respectively, to prune networks

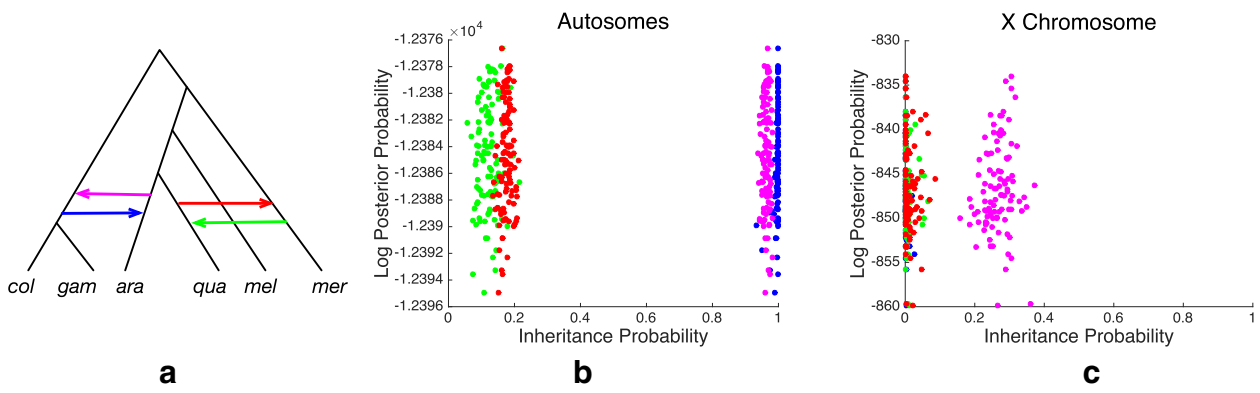

Fig. 6 Analysis of a mosquito data set. **a** The species network tested in our method. **b** The distribution of inheritance probabilities returned by our method using bootstrap gene trees from autosomes. **c** The distribution of inheritance probabilities returned by our method using bootstrap gene trees from X chromosome. The colors in panels (**b**) and (**c**) correspond to the reticulation edge colors in panel (**a**)

and *An. coluzzii* (indicated by blue and pink reticulation edges) and gene flow from *An.merus* to *An. quadriannulatus* (indicated by green reticulation edge). Later, Wen et al. [23] reanalyzed this data set and reported a different species network which is the network in Fig. 6a excluding the green reticulation edge. It was inferred by adding reticulations on the underlying species tree of [12] under maximum likelihood using bootstrap gene trees from the autosomes. The difference between these two hypothesis is the direction of gene flow between *An. quadriannulatus* and *An. merus*.

We reanalyze this data set using our method, mainly focusing on testing the gene flow between *An. quadriannulatus* and *An. merus* and the other two reticulations that both [12] and [23] agreed on. We used the gene trees of [23], which were reconstructed from 2791 loci sampled at least 64 kb apart from autosomes, including 669 from 2L, 849 from 2R, 564 from 3L and 709 from 3R. For every locus, 100 bootstrap trees were built. Then Eq. (1) becomes

$$p(N,\theta|\mathcal{G}) \propto p(N,\theta) \prod_{G\in\mathcal{G}} \frac{\sum_{g\in G} p(g|N,\theta)}{|G|} \qquad (3)$$

where G contains all bootstrap gene trees from a given locus. The method took close to 2 days to obtain the results. Figure 6b shows the posterior of the inheritance probability samples computed by the Gibbs sampler. As the figure shows, for the pink and blue reticulation edges, which [12] and [23] agreed on, the inheritance probabilities are very close to 1, which suggests that the data support an underlying "backbone" tree that groups (col, gam) with *ara*, in agreement with the tree inferred by maximum likelihood in [23]. As for the red and green reticulation edges, the posterior samples indicate non-negligible amount of introgression along both of these edges.

We repeated the analysis using gene tree data from the X chromosome. This data set contains 228 loci sampled at least 64 kb apart from X chromosome and 100 bootstrap trees were built for each locus. The posterior samples of the four inheritance probabilities are shown in Fig. 6c. The inheritance probabilities of the blue, red and green reticulation edges are all close to 0, which makes sense given that the species tree in [12] was inferred based on the X chromosome data. For the pink reticulation edge, the inheritance probabilities are between 0.2 and 0.4, which indicates that there is introgression from *An. arabiensis* to the ancestor of *An. gambiae* and *An. coluzzii* on X chromosome, in agreement with [23].

Analysis of a modern bird data set

We reanalyzed the modern bird data set of [13]. The original data set contains 48 species representing all orders of Neoaves. In the species tree the authors reported, the three vocal learners (*Hummingbirds*, *Parrots* and *Oscines*) are not monophyletic. *Hummingbirds*, in particular, were placed far from the other two. An interesting question in this context is whether there was convergent evolution in vocal learning or it was shared among these three species via introgression. To investigate this question, we first pruned the data set from 48 species to 16 for computational feasibility. We selected *Medium Ground-Finch* to represent *Oscines*, *Budgerigar* to represent *Parrots*, and then we arbitrarily selected one species from each of the well-supported clades. Lastly, we added reticulation edges between every among *Oscines*, *Parrots* and *Hummingbirds*. The resulting species network is shown in Fig. 7a. We downloaded the maximum likelihood gene trees of [13], including 8251 based on exons, 2516 based on introns and 3679 based on ultra-conserved elements. We used *Struthioniformes* (*Ostrich*) or *Tinamiformes* (*Tinamous*) to root all the gene trees. For gene trees that do not contain either of these two, we excluded them

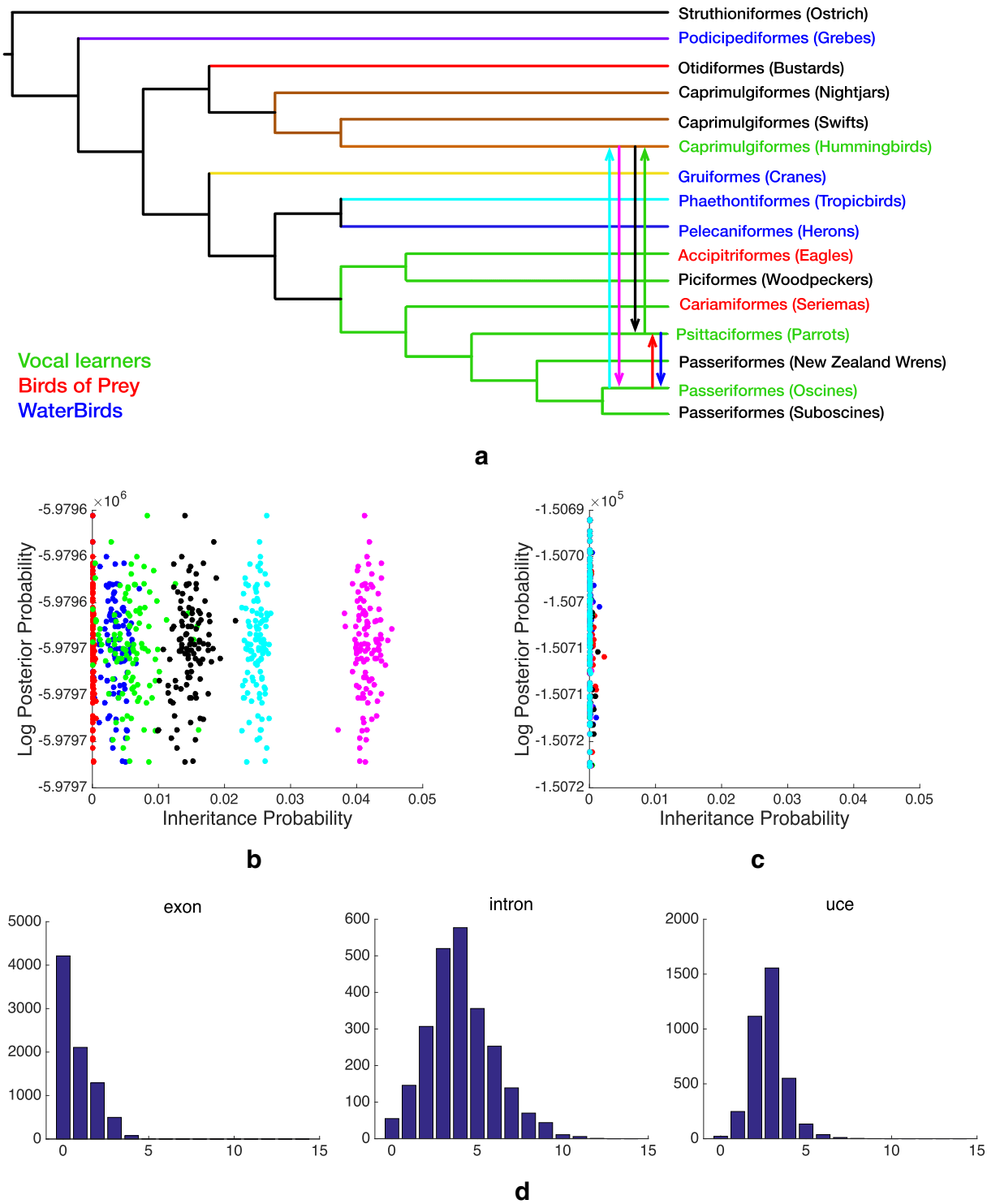

Fig. 7 Analysis of a modern bird data set. **a** The species network explored by our method, which contains six pairwise reticulation edges among the three vocal learners. **b** The posterior of the inheritance probability samples associated with the six reticulation edges using all 14,357 gene trees from exons, introns and ultraconserved elements. **c** The posterior of the inheritance probability samples associated with the six reticulation edges using 524 gene trees (from introns) that have more than 5 internal branches with at least 70 bootstrap support. **d** The number of gene trees with high bootstrap support. The number of internal branches in the gene trees with at least 70 bootstrap support is on the x-axis. The number of gene trees that have a given number of branches with at least 70 bootstrap support is on the y-axis. The three panels from *left to right* correspond to gene trees from exons, introns and ultra-conserved elements, respectively

from our analysis. We ended up with a total of 14,357 gene trees.

Since the data set is too large for full-likelihood calculations, we used pseudo-likelihood [18]. The method took close to 5 days to obtain the results Fig. 7b shows the posterior of the inheritance probabilities collected from the Gibbs sampler when using the entire gene tree data set. The results indicate non-negligible gene flow between *Parrots* and *Hummingbirds* (in cyan and pink) and from *Hummingbirds* to *Parrots* (in black), but negligible inheritance probabilities (and, consequently, gene flow) between *Parrots* and *Oscines* (in red and blue) and from *Parrots* to *Hummingbirds* (in green). However, given that a large majority of the gene trees of [13] have poor bootstrap support, the question becomes: Is this detected introgression real or an artifact of the poor support of gene trees (errors in gene trees can masquerade as introgression signal). Figure 7d provides a clear picture of how little resolution the gene trees of [13] had: The great majority of trees inferred from exons and ultra-conserved elements had fewer than 5 internal branches with support exceeding 70. Therefore, we repeated the analysis only using gene trees that have at least 6 internal branches with bootstrap support of at least 70. This data set consists of 524 gene trees only. When we used this gene tree data set, the results were negligible inheritance probabilities along all six reticulation edges (Fig. 7c). In other words, the gene trees with strong signal support a treelike evolutionary hypothesis of this group of birds, indicating the possibility that vocal learning has undergone convergence in this group, at least as supported by this data. This further attests the strength of our method: While it uses networks for evolutionary exploration, it returns treelike hypotheses when they are supported by the data.

Conclusions

In this paper, we showed how to use Gibbs sampling to explore phylogenetic hypotheses over evolutionary phylogenetic networks. These hypotheses could involve reticulate and non-reticulate evolutionary processes simultaneously. We showed how pseudo-likelihood could be used to speed up the computation and make the analysis of large data sets feasible. We demonstrated the power of our method to explore phylogenetic hypotheses on two biological data sets, and assessed its performance on simulated data. An open-source implementation of the method is publicly available as one of the functionalities in the PhyloNet software package [11].

The analysis of the modern bird data set highlights a very important issue that is relevant not only to network analysis, but to all phylogenetic analyses, namely, the effect of error in gene tree estimates on methods that use those estimates as the primary data for inference. When all gene trees in the data set were used, regardless of their support, large extents of introgression were estimated. However, when only well-supported gene trees were used, introgression patterns mostly disappeared. Gene tree topological estimation errors masquerade as signal for biological causes of incongruence. In our case, these causes could be incomplete lineage sorting or introgression. Therefore, to avoid erroneous inferences, particularly false positives, it is very important that only well-supported gene tree topologies are used in the analyses.

The work of [19] is most relevant to the method presented here. In [19], the phylogenetic network topology and its associated parameters are all sampled, which gives rise to mathematical and computational challenges arising from quantifying convergence and summarizing phylogenetic network topologies. Nevertheless, the method is powerful in sampling the posterior of phylogenetic networks and associated parameters, and is useful when that posterior is the quantity of interest. Our proposed method here differs in that we see its primary use in sampling the posterior of only the continuous parameters (branch lengths and inheritance probabilities) of a given set of phylogenetic network topologies that reflect evolutionary hypotheses of interest. Since the network topology is fixed during the sampling, summarizing the sampled values of the continuous parameters is straightforward in our proposed method.

It is important to note that while we illustrated our method on evolutionary hypotheses formed by adding horizontal edges to a given (species) tree, the method treats the phylogenetic topology as a network and does not designate any trees inside the network as a species tree. Furthermore, since the network is fixed during the sampling, any evolutionary relationship that is captured by the analyzed network cannot be uncovered (which is another difference between this method and that of [19]).

While we focused here on the multispecies network coalescent [10, 19], statistical models that incorporate, for example, gene duplication and loss, could be added naturally to the framework.

Declarations

Publication of this article was funded in part by grants DBI-1062463, CCF-1302179, OCI-0959097 (Data Analysis and Visualization Cyberinfrastructure), and CNS-1338099 (Big-Data Private-Cloud Research Cyberinfrastructure) from the National Science Foundation of the United States of America.

This article has been published as part of *BMC Genomics* Vol 17 Suppl 10, 2016: Proceedings of the 14th Annual Research in Computational Molecular Biology (RECOMB) Comparative Genomics Satellite Workshop: genomics. The full contents of the supplement are available online at https://bmcgenomics.biomedcentral.com/articles/supplements/volume-17-supplement-10.

Availability of data and materials

Not applicable.

Authors' contributions
YY, CJ and LN conceived of the study and designed it. YY implemented the method and conducted all the data analyses. All authors wrote the manuscript. All authors have read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Published: 11 November 2016

References

1. Edwards SV, Xi Z, Janke A, Faircloth BC, McCormack JE, Glenn TC, Zhong B, Wu S, Lemmon EM, Lemmon AR, Leache AD, Liu L, David CC. Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. Mol Phylogenet Evol. 2016;94:447–62.
2. Springer MS, Gatesy J. The gene tree delusion. Mol Phylogenet Evol. 2016;94:1–33.
3. Nakhleh L. Evolutionary phylogenetic networks: models and issues In: Heath L, Ramakrishnan N, editors. The Problem Solving Handbook for Computational Biology and Bioinformatics. New York: Springer; 2010. p. 125–58.
4. Bapteste E, van Iersel L, Janke A, Kelchner S, Kelk S, McInerney JO, Morrison DA, Nakhleh L, Steel M, Stougie L, Whitefield J. Networks: expanding evolutionary thinking. Trends Genet. 2013;29(8):439–41.
5. Nakhleh L. Computational approaches to species phylogeny inference and gene tree reconciliation. Trends Ecol Evol. 2013;28(12):719–28.
6. Huson DH, Rupp R, Scornavacca C. Phylogenetic Networks: Concepts, Algorithms and Applications. New York: Cambridge University Press; 2010.
7. Morrison DA. Networks in phylogenetic analysis: new tools for population biology. Int J Parasitol. 2005;35(5):567–82.
8. Huson DH. SplitsTree: a program for analyzing and visualizing evolutionary data. Bioinformatics. 1998;14(1):68–73.
9. Yu Y, Degnan JH, Nakhleh L. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. PLoS Genet. 2012;8:1002660.
10. Yu Y, Dong J, Liu K, Nakhleh L. Maximum likelihood inference of reticulate evolutionary histories. Proc Natl Acad Sci. 2014;111: 16448–53.
11. Than C, Ruths D, Nakhleh L. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinforma. 2008;9(1):322.
12. Fontaine MC, Pease JB, Steele A, Waterhouse RM, Neafsey DE, Sharakhov IV, Jiang X, Hall AB, Catteruccia F, Kakani E, et al. Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science. 2015;347(6217):1258524.
13. Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SY, Faircloth BC, Nabholz B, Howard JT, et al. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science. 2014;346(6215): 1320–31.
14. Yu Y, Ristic N, Nakhleh L. Fast algorithms and heuristics for phylogenomics under ils and hybridization. BMC Bioinforma. 2013; 14:6.
15. DeGiorgio M, Degnan JH. Robustness to divergence time underestimation when inferring species trees from estimated gene trees. Syst Biol. 2014;63(1):66–82.
16. Geman S, Geman D. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. IEEE Trans Pattern Anal Mach Intell. 1984;6:721–41.
17. Felsenstein J. Inferring Phylogenies. Sunderland: Sinauer Associates, Inc.; 2003.
18. Yu Y, Nakhleh L. A maximum pseudo-likelihood approach for phylogenetic networks. BMC Genomics. 2015;16:10.
19. Wen D, Yu Y, Nakhleh L. Bayesian inference of reticulate phylogenies under the multispecies network coalescent. PLoS Genet. 2016;12(5): 1006006.
20. Hudson RR. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics. 2002;18:337–8.
21. Rambaut A, Grassly NC. Seq-gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comp Appl Biosci. 1997;13:235–8.
22. Stamatakis A. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9): 1312–13.
23. Wen D, Yu Y, Hahn MW, Nakhleh L. Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. Mol Ecol. 2016;25:2361–372.