Unifying Gene Duplication, Loss, and Coalescence on Phylogenetic Networks

Peng Du, Huw A. Ogilvie, and Luay Nakhleh(✉)

Rice University, Houston, TX 77005, USA
{peng.du,huw.a.ogilvie,nakhleh}@rice.edu

Abstract. Statistical methods were recently introduced for inferring phylogenetic networks under the multispecies network coalescent, thus accounting for both reticulation and incomplete lineage sorting. Two evolutionary processes that are ubiquitous across all three domains of life, but are not accounted for by those methods, are gene duplication and loss (GDL).

In this work, we devise a three-piece model—phylogenetic network, locus network, and gene tree—that unifies all the aforementioned processes into a single model of how genes evolve in the presence of ILS, GDL, and introgression within the branches of a phylogenetic network. To illustrate the power of this model, we develop an algorithm for estimating the parameters of a phylogenetic network topology under this unified model.

We demonstrate the application of the model and the accuracy of the algorithm on simulated as well as biological data.

Our work adds to the biologist's toolbox of methods for phylogenomic inference by accounting for more complex evolutionary processes.

Keywords: Phylogenetic network · Coalescence · Introgression · Gene duplication and loss

1 Introduction

Independently evolving lineages of eukaryotic organisms are typically referred to as *species* (they may also be referred to as *populations* depending on the context and operational definition of those terms). Over evolutionary time scales, species lineages bifurcate to form two descendant species from a single ancestral species. This gives rise to a *species tree*, which is a phylogenetic tree describing the evolutionary history of a set of species.

Estimating a species tree is challenging as gene trees are expected to be discordant with the species tree because of several well known processes. The first

This work was supported in part by NSF grants DBI-1355998, CCF-1302179, CCF-1514177, CCF-1800723, and DMS-1547433.

Z. Cai et al. (Eds.): ISBRA 2019, LNBI 11490, pp. 40–51, 2019.
https://doi.org/10.1007/978-3-030-20242-2_4

process leading to discordance is incomplete lineage sorting (ILS), where multiple versions or *alleles* of a gene persist in a species up through to its ancestral species [9]. The second is horizontal gene transfer (HGT) through hybrid speciation [11], introgression [10], and speciation with gene flow [13]. This can lead to gene coalescent times which are younger than the earliest speciation event separating the corresponding species. The third is gene duplication and loss (GDL), where new copies of a gene are created at new loci in the genome, so that the relationship between sequences from different species at different loci (paralogs) reflects the duplication and loss process rather than the speciation process [4].

ILS has been addressed by years of research into the multispecies coalescent (MSC), a mathematical model which describes the evolution of gene trees within a species tree and naturally accommodates ILS [4]. In the MSC, the relationship between sequences from different species at orthologous (as opposed to paralogous) loci is represented by a gene tree, evolving within a species tree, and constrained so that its coalescent times must be older than the corresponding most recent common ancestors (MRCAs).

More recently, HGT has been addressed by generalizing the MSC model to the multispecies network coalescent (MSNC) model, which represents the evolutionary history of species as a phylogenetic network [22]. This flexible model of reticulate evolution can naturally accommodate hybrid speciation [25] and introgression [21]. Implementations of this method include `mcmc_seq` in PhyloNet [20,23] and SpeciesNetwork in BEAST [25].

GDL has been addressed by the development of models which add a third layer to the MSC between the species tree and the gene trees. This is known as the locus tree, and it contains vertices encoding duplication events, as well as vertices which directly correspond to the speciation vertices of the species tree [16]. The duplicate copy of a gene is assumed to reside in a new unlinked locus, so that there are multiple copies of a gene present in a single genome. The leaves of a single locus tree can therefore represent multiple loci, and the source of data in this model may be more appropriately termed "gene families" (cf. "genes").

DLCoal, the original implementation of the three-layer model [16], is relatively inflexible. It takes as input a gene tree topology, a species tree fixed in topology, branch lengths and effective population sizes, and rates of gene duplication and loss. From such input data it can estimate the locus tree, the mapping of gene tree coalescent vertices to locus tree branches, and the mapping of speciation vertices in the locus tree to the species tree. DLCoal also relies on the accuracy of the supplied gene tree topology, which may contain errors due to the gene tree inference method or insufficient information in the original multiple sequence alignment (MSA). A later method, DLC-Coestimation [24], avoids that potential issue by jointly estimating the gene tree along with the locus tree and reconciliations and mapping directly from a gene family MSA.

The most recent implementation of the three-layer model jointly estimates the species, locus and gene tree topology and times, as well as general parameters including duplication and loss rates from the MSAs of multiple gene families [5]. In a simulation study, this method was able to successfully infer the species tree topology, and outperformed using the MSC model alone (without accounting for GDL) when estimating species divergence times [5].

While the above methods either account for both ILS and HGT, or for both ILS and GDL, no model has been designed or implemented that accounts for all three processes which generate gene tree discordance; ILS, HGT and GDL. Here we present a new model which extends the MSNC to a three-layer model by adding a locus network between the species network and gene trees. This new model accounts for HGT at the species network level, GDL at the locus network level, and ILS at the gene tree level. We have implemented a maximum *a posteriori* (MAP) search for this model which jointly estimates the speciation times, inheritance probabilities and duplication and loss rates. Using simulation experiments, we show that it can accurately infer the aforementioned parameters.

We also used simulated data and an empirical data set of six yeast species to study the difference in accuracy between our new method and an MSNC method which does not account for GDL. Results from those experiments showed that accounting for GDL in addition to ILS and HGT is particularly important when estimating reticulation times.

2 Methods

Similar to the three-layer model of [16], we develop a three-layer model that uses a locus network (different from the locus tree of [16]) as an intermediate layer between the species network and gene tree. This structure allows for unified modeling of coalescence and GDL, where all coalescence events are captured by the relationship between the gene tree and locus network, and all GDL events are captured by the relationship between the locus network and phylogenetic network. The reticulation events (e.g., introgression) are captured by the fact that the species and locus structures are both networks, rather than trees.

2.1 The Three-Layer Model

A species network $\mathbb{S} = (V(S), E(S), \tau^S)$ is a directed acyclic graph depicting the reticulate evolutionary histories of a set of species where $V(S)$ is the set of vertices in the network, $E(S)$ is the set of edges and τ^S contains the set of branch lengths of the edges. We use S to denote $\{V(S), E(S)\}$. Further, $V = r \cup V_L \cup V_T \cup V_N$ where r is the root of the network, V_L is the set of leaf vertices, V_T denotes the set of tree vertices with two children and one parent and V_N represents the set of reticulation vertices with one child and two parents. The set of all internal vertices is $IV(S) = r \cup V_T \cup V_N$. If vertex u has only one parent, we call this parent $pa(u)$. The set of children of u is denoted as $c(u)$. For each reticulation vertex u with two parents v and w, there is an inheritance probability $\gamma \in [0, 1]$ such that the probability of locus u inheriting from v is γ and inheriting from w is $1 - \gamma$. $\mathit{\Gamma}$ is a vector of all inheritance probabilities for all vertices in V_N, $\mathit{\Gamma}((v, u)) = \gamma$ and $\mathit{\Gamma}((w, v)) = 1 - \gamma$. The population sizes are denoted as N^S and the population size on branch $e(u, v)$ is $N^S((u, v))$.

Locus Networks and Locus-Network-to-Species-Network Reconciliation. A locus network $\mathbb{L} = (V(L), E(L), \tau^L)$ is generated by applying duplication and loss events onto the species network with a top-down birth-death process [1,2,18]. Birth events create new loci by duplicating an existing locus, and death (loss) events eliminate loci so that it will have no sampled descendants. Fully describing the result of this process requires a reconciliation R^L from the locus network to the species network, where the vertices on the locus network can be mapped to either the vertices or the branches of the species network. If $u \in V(L)$ is mapped to a species network vertex, then we call it a speciation vertex; the set of speciation vertices is denoted as $V_S(L)$. If it is mapped to a species network branch; we call it a duplication vertex and the set of duplication vertices is denoted as $V_D(L)$. Branches with no existing leaf vertices are pruned out (Fig. 1). For a duplication, a new locus is generated, so a mapping $\delta(u, v) = 1$ or $\delta(u, v) = 0$ is used to indicate whether (u, v) leads to the new (daughter) locus or if (u, v) is the mother branch where u is the duplication vertex. The population size of branch $e = (u, v)$ in the locus network is the population size of the branch $e' = (w, x)$ on the species network where $R^L(u) = (w, x)$ or $R^L(v) = (w, x)$ or $R^L(u) = w, R^L(v) = x$. Similarly, $\Gamma((u, v)) = \Gamma((w, x))$ where $(w, x) \in E(S)$ if $R^L(u) = (w, x)$ or $R^L(v) = (w, x)$ or $R^L(u) = w, R^L(v) = x$. It is important to note that reticulation edges present in the species network may be deleted from the locus network (as in the (F, X) branch leading to the B2 locus in Fig. 1), so the locus network can be a tree or more tree-like with fewer reticulation vertices than the species network.

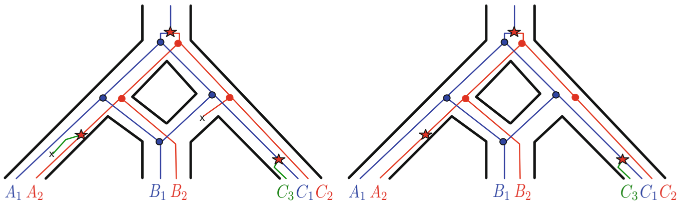


Fig. 1. A gene duplication and loss scenario inside of a species network on three species A, B, and C. (Left) The complete locus network embedded in the species network, produced by a birth-death process, and containing all duplication and loss events. (Right) Lineages in the locus network with no sampled loci due to loss events are pruned from the locus network, resulting in the observed locus network. Extinct lineages are deleted. Duplication, loss, and speciation/hybridization events are represented by $\star$, $\times$, and $\bullet$, respectively. New lineages arising from duplication are colored red and green. (Color figure online)

Gene Trees and Gene-Tree-to-Locus-Network Reconciliation. A gene tree $\mathbb{G} = (V(G), E(G), \tau^G)$ describes the evolution of lineages and the definitions of vertices are similar with those in the species network and locus network. The reconciliation from the gene tree to the locus network is denoted by R^G. The two reconciliations R^L and R^G are collectively denoted by R. For each locus network branch $e = (u, w)$ with $\delta((u, w)) = 1$, the coalescent time of every gene vertex mapped to the leaf vertices under w must be more recent than u. Also, we define M as the mapping from the gene tree leaf vertex-set to the locus network leaf vertex-set. M indicates what gene is from what locus in the locus network. Figure 2 shows the reconciliations from the gene tree to the locus network (R^G) and from the locus network to the species network (R^L).

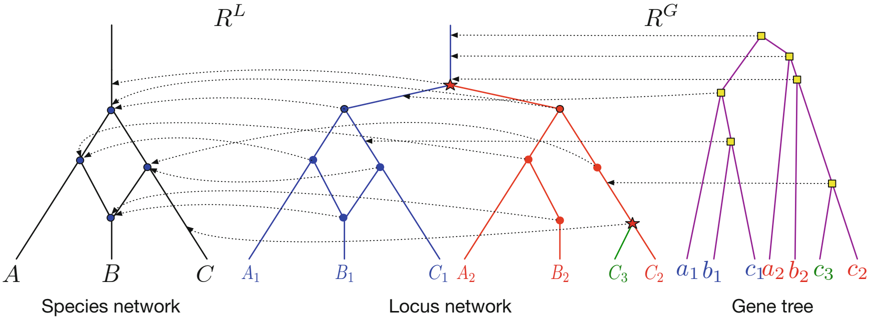


Fig. 2. Gene duplication/loss events are obtained by mapping the nodes of the locus network onto the branches of the species network, via reconciliation R^L (the dotted arrows from the locus network to the species network). Coalescence events are obtained by mapping the gene tree nodes onto the branches of the locus network, via reconciliation R^G (the dotted arrows from the gene tree to the locus network). Duplication, loss, speciation/hybridization, and coalescence events are represented by $\star$, $\times$, $\bullet$, and $\blacksquare$, respectively.

2.2 Model Assumptions

In this model, we need to make some assumptions as made in [5,16,24].

1. After the duplication, the daughter locus becomes totally unlinked and any further evolution of the mother and daughter loci, as well as the coalescent histories of the mother and daughter genes, are independent conditional on the species network topology, times and population sizes. Thus we can calculate the coalescent probabilities separately for each locus, and use the product as the gene family coalescent probability.
2. At a locus level, hemiplasy [3] is assumed to be non existent in this model. In other words, for each duplication and loss event, the resulting addition or deletion of locus will be transmitted universally to all descendent species. This allows us to explain all unobserved loci by means of gene loss.
3. In our present implementation, one individual per species is sampled for each locus.

2.3 Probability Distribution

For a species network $\mathbb{S}$ and a set of gene families $\mathbb{GF}$ with each member $\mathbb{GF}_i = (\mathbb{L}_i, \mathbb{G}_i, R_i, M_i, \delta_i^L)$, and parameters θ, the posterior $p(\mathbb{S}, \mathbb{GF}, \theta | D)$ given observed DNA sequences D is

$$p(\mathbb{S}, \mathbb{GF}, \theta | D) \propto \prod_i p(\mathbb{GF}_i | \mathbb{S}, \theta) \times p(D_i | \mathbb{GF}_i) \times p(\theta)$$

where D_i is the DNA sequences for $\mathbb{GF}_i$ and $\theta = \{\mu, \lambda, \Gamma, N^S\}$ which are the duplication rate, loss rate, substitution rate, inheritance probabilities and population size respectively. The term $p(\mathbb{GF}_i|\mathbb{S}, \theta)$ can be decomposed into (we drop the subscript i for readability) the product $p(G, \tau^G, R^G | L, \tau^L, \delta^L, M, \Gamma, N^S) \times p(M|L, \tau^L, R^L, \delta^L) \times p(L, \tau^L, R^L, \delta^L | S, \tau^S, \mu, \lambda)$, and we have $p(D|\mathbb{GF}_i) = p(D|G, \tau^G)$. The term $p(G, \tau^G, R^G | L, \tau^L, \delta^L, M, \Gamma, N^S)$ is the probability of the gene tree coalescing in the locus network under a bounded coalescence model where gene lineages originated from gene duplication events must coalesce earlier than the duplication event. The bounded coalescence model is extended from [16] and gains the capacity to handle hybridization events. The details are in [6].

The term $p(M|L, \tau^L, R^L, \delta^L)$ is the probability of the map of gene tree leaves to locus network leaves. Since we assume no prior knowledge of locus information of each sampled gene copy from a certain species, the mapping has a uniform distribution based on the number of possible permutations:

$$p(M|L, \tau^L, R^L, \delta^L) = \prod_{x \in L(S)} \frac{1}{|u : R^L(u) = x|!}. \tag{1}$$

The number of permutations is constant for a given data set D, so for identification of the MAP topology or algorithms like MCMC which use unnormalized posterior probabilities not scaled by $1/P(D)$, the calculation of this prior is unnecessary. The term $p(L, \tau^L, R^L, \delta^L | S, \tau^S, \mu, \lambda)$ is the probability of the locus network generated inside of the species network with duplication rate μ and loss rate λ and is also derived in [6]. The term $p(\mathbb{S})$ is the prior of the species network which is a compound prior with uniform prior on the topology and exponential prior on divergence times as in [7,19].

2.4 MAP Inference of the Parameters of a Fixed Network Topology

Our goal is to find the maximum *a posteriori* (MAP) estimate of the parameters; that is,

$$(\mathbb{S}^*, \mathbb{GF}^*, \theta^*) = \text{argmax}_{(\mathbb{S}, \mathbb{GF}, \theta)} p(\mathbb{S}, \mathbb{GF}, \theta | D). \tag{2}$$

In this present work, we will focus on inferring the species network parameters—times, population sizes and inheritance probabilities—with the topology being fixed, as well as locus networks, gene trees and reconciliations between them. General parameters such as duplication and loss rates are also inferred. Because of the hierarchical nature of the generative model, changes on higher level components will influence lower level components as well. For example, changing the heights of the species network vertices will also change the heights of corresponding locus network vertices. We developed four groups of operators, each working on different levels of the model, which we describe in detail in [6]. The first group makes changes to the species network and can also alter the locus networks and gene trees. The second group changes the locus network and can also alter the gene trees. The third group makes changes to the gene trees alone, while the fourth applies to the macroevolutionary rates, which in our implementation is limited to the duplication and loss rates.

2.5 Results and Discussion

2.6 Performance on Simulated Data

Simulation Setup. We simulated DNA sequence data for multiple gene families with our gene tree simulator and Seq-Gen [14]. Our gene tree simulator employs the hybridization-duplication-loss-coalescence model and operates in two phases. First, it generates the locus network within a predefined species network by simulating duplications and losses.

Then the gene tree is simulated under a coalescence model along the locus network. If the gene lineages could not coalesce before the duplication event backward in time, it will be rejected and retried until it coalesces after the event, up to 10^8 attempts, beyond which the locus network will be rejected and regenerated. Locus networks with fewer than 3 extant species will be rejected. Once the gene trees were generated, the program Seq-Gen [14] was used to simulate the evolution of DNA sequences down the gene trees under a specified model of evolution. In all simulations reported here, we used the Jukes-Cantor model of evolution [8] to generate 1000 bp long DNA sequences. For Experiments 1 to 3, we used the network of Fig. 3 as the model species network. Population sizes are given as the number of diploid individuals, and specified duplication/loss rates and population sizes were set to be the same across all branches of the model networks. A mutation rate of 10^{-9} was used for all simulation experiments.

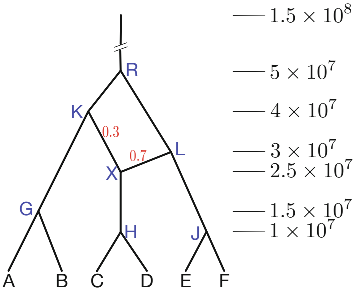


Fig. 3. The model network used to simulate data for Experiments 1 to 3. The values on the right correspond to divergence times of the nodes in number of generations. The inheritance probability values are shown on the reticulation edges.

Experiment 1: Testing the Effects of GDL Rate and Population Size. In this experiment, different settings of duplication/loss rates and population sizes were used to test how these parameters would affect the accuracy of inferences. The duplication and loss rates (both were equal) used were 5×10^{-10}, 10^{-9} and 2.5×10^{-9} and the population sizes were 10^6, 4×10^6, and 8×10^6. For each of the 9 different settings of duplication/loss rates and population size, we generated 10 replica each with 50 gene families and ran 15 million iterations for each data set. First, we calculated the average difference of the estimated divergence times and the true values in population mutation rate units for the 9 settings based on the 10 replica for each setting. Most estimates of the divergence times are accurate across different settings, with the exception of the reticulation time "X" which appears to be less identifiable at smaller population sizes (see [6]). We calculated the difference between the estimated inheritance probabilities and the true value (0.3) on (K, X) (see [6]). No consistent trend was observed in the accuracy of inheritance probability estimates over the ranges of population size and duplication and loss rates studied. To assess the accuracy of locus networks and gene trees, we calculated the topological error in metrics developed by [12] between the estimated and true locus networks and RF distance between true and estimated gene trees [17]. Overall our method shows very good accuracy (indicated by topological distances close to 0). The average distance for the locus networks increases as the duplication and loss rate increases, but it appears invariant to varying population size. This makes sense because the locus networks are determined by the duplication and loss events not the ILS events. If ILS, GDL and HGT are absent all gene tree topologies will identical to the species tree topology, and be perfectly accurate when the species tree topology is fixed at the truth. However in our model gene trees can vary because of all three processes. The prevalence of ILS is partly dependent on population sizes, and therefore it is unsurprising that we show gene tree topological error consistently increasing as population sizes get larger (see [6]). Finally, we assessed the method's performance in terms of estimating the duplication and loss rates. As the results show, the method performs well at estimating both rates under the range of population sizes and duplication and loss rates studied (see [6]).

Experiment 2: Testing the Effect of the Number of Gene Families. In order to determine how our method performs given larger or smaller data sets, we varied the number of gene families (5, 10, 25, and 50) under one setting of duplication/loss rate (2.5×10^{-9}) and population size (4×10^{6}). 10 replica for each number of gene families were simulated and 15 million iterations were run for each data set. Results (see [6]) show that even for 5 gene families, a relatively small number, the estimated divergence times are generally accurate especially for H and R. The accuracy and precision improve as more gene families are used for example for nodes G, J and L.

We tested the inference of other parameters. Figure 4(a) shows that both the accuracy and precision of the inheritance probability improved for larger numbers of gene families. The accuracy of the duplication rate both appear to improve slightly with more data. The loss rate, while accurately estimated, did not show any consistent trends. As the results show, the accuracy of the locus networks and of the gene trees seems to be stable across different settings in terms of both mean and standard deviation. As gene tree topologies, while independent, are conditioned on the species network topology, when the species network topology is fixed even without any data there will already be a lot of information in the model on the gene tree topologies. Also, the locus networks are independent for each gene tree conditioned on the species network topology. So increasing the number of gene trees will not improve the overall accuracy of gene tree estimates to the same extant as when jointly estimated with the species tree or network topology (Fig. 4(b)).

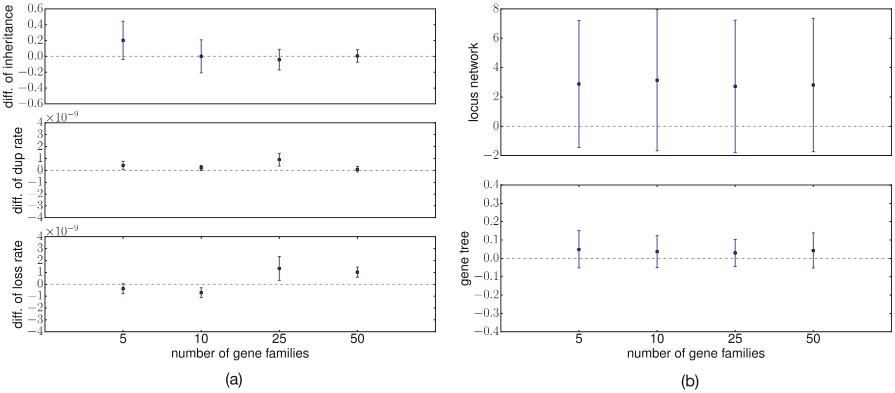


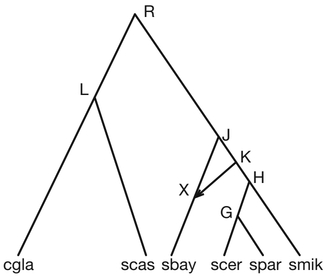
Fig. 4. (a) The difference of estimated parameters from the true values. Top: difference between estimated inheritance probability and the true value (0.3) on (H, X). Middle: difference of estimated duplication rate and true value. Bottom: difference of estimated loss rate and true value. The number of gene families used as input to the inference method is shown on the x-axis. Standard deviations are represented by vertical bar. (b) The average topological distances between the inferred and true networks or trees. Top: Locus network difference. Bottom: Gene tree difference. The number of gene families used as input to the inference method is shown on the x-axis. Standard deviation is represented as vertical bar.

Experiment 3: Comparing Inference With and Without GDL. In this experiment we set out to test how a method that accounts only for incomplete lineage sorting but ignores duplication and loss would perform as compared to our model here. To achieve this, we ran our method and a Bayesian MCMC species network inference method (the `mcmc_seq` command, with the species network fixed, and using the MAP estimation) in PhyloNet [23] which implements the method of [20]. We simulated 10 replica under duplication and loss rates of 2.5×10^{-9} and population size 10^7 and 50 gene families for each data set. For each gene family we randomly selected one gene copy for each species if there was at least one. As a result, around half of the sequences in the gene families were kept after this pruning of the data sets. We fed the sequences to both methods and ran them both for 15 million iterations. Our results show that our method, which accounts for gene duplication and loss even with a single sampled locus per species, more accurately estimated speciation and reticulation times. This was particularly true of the reticulation vertex, where `mcmc_seq` dramatically underestimated the reticulation time (see [6]). Also, we have a better estimation of the inheritance probabilities than mcmc_seq. Our estimation is 0.268 and mcmc_seq had estimation of 0.464 where the true value is 0.3.

2.7 Biological Data

We used the yeast genome data set with duplications reported on http://compbio.mit.edu/dlcoal/ and randomly selected two data sets restricted to six genomes. One consists of 100 gene families each with exactly one copy for each species with alignments 1000nt–2000nt in length; the other consists of 100 gene families with possibly multiple or 0 copies for each species with alignments 1000nt–2000nt in length. We used 10^{-10} as duplication and loss rate and 4×10^{-10} as mutation rate and 10^7 as population size which are comparable with the settings used in [15]. Then we fed mcmc_seq with the first data set and ran the command 10 times for 15 million iterations each with the maximum number of reticulation vertex set to be one. The most prevalent topology is shown in Fig. 5(a) and appeared in 7 of the 10 runs. We then fed our method with the second data set and run 7 times each with 15 million iterations. A table of the average estimated divergence times is given in Fig. 5(b). We can see that most of the divergence times are similar and the only significant differences are at the divergence times of vertices J, X and L. Given our method is better at estimating divergence times given results from Experiment 3, it seems that the ones obtained by our method here are probably more accurate estimations.

The inheritance probability on branch (R, X) estimated by mcmc_seq was 0.503 ± 0.147 while the value estimated by our method was 0.461 ± 0.091. The error is the standard deviation among runs, and shows that the estimated inheritance probabilities of the two methods are very close.



(a)

	Our method	mcmc_seq
G	0.04068 ± 0.001	0.0401 ± 0.001
H	0.0708 ± 0.001	0.0677 ± 0.002
X	0.02851 ± 0.015	0.0626 ± 0.0292
K	0.0937 ± 0.003	0.0911 ± 0.004
J	0.1187 ± 0.005	0.1227 ± 0.004
L	0.2321 ± 0.002	0.1877 ± 0.007
R	0.2415 ± 0.002	0.1955 ± 0.008

(b)

Fig. 5. (a) The yeast species network topology inferred by mcmc_seq on the 100 gene families. (b) Mean and standard deviation of the estimated divergence times of mcmc_seq and our method (std's smaller than 0.001 are rounded to 0.001).

3 Conclusions

In this work, we developed a probabilistic model that simultaneously accounts for hybridization, gene duplication, loss and ILS. We also devised a stochastic search algorithm for parameterizing phylogenetic networks based on this model. This algorithm provides estimates of evolutionary parameters, as well as gene histories and their reconciliations. Results based on simulation studies show good performance of the algorithm as well as insights obtained by employing the new model as compared with existing models that exclude gene duplication and loss.

We identify three natural directions for future research. First, while in this work we assumed a fixed phylogenetic network topology, in most empirical studies such a topology is not given or known. Developing a method that infers the phylogenetic network, along with all the parameters that the current method estimates, is essential for proper application of the model. Second, while this work focused on obtaining point estimates of the phylogenetic network's parameters, developing a method that estimates a posterior distribution on the space of phylogenetic networks and their parameters would provide additional information, including assessment of statistical significance and the uniqueness and distinguishability of optimal solutions. Third, the computational bottleneck in this domain stems from the time it takes to compute the likelihood of a given point in the parameter space as well as from the need to walk an enormous and complex space of such parameters. For example, it took between 15 and 20 h for a single run of 15 million iterations on a data set with four or five species and 50 gene families. Developing algorithmic techniques and potentially alternative likelihood functions to speed up these calculations is imperative for this work to be applicable to data sets of the scale that biologists can now generate using the latest sequencing technologies.

References

1. Åkerborg, Ö., Sennblad, B., Arvestad, L., Lagergren, J.: Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. Proc. Natl. Acad. Sci. **106**(14), 5714–5719 (2009)
2. Arvestad, L., Lagergren, J., Sennblad, B.: The gene evolution model and computing its associated probabilities. J. ACM (JACM) **56**(2), 7 (2009)
3. Avise, J.C., Robinson, T.J.: Hemiplasy: a new term in the lexicon of phylogenetics. Syst. Biol. **57**(3), 503–507 (2008)
4. Degnan, J.H., Rosenberg, N.A.: Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. **24**(6), 332–340 (2009)
5. Du, P., Nakhleh, L.: Species tree and reconciliation estimation under a duplication-loss-coalescence model. In: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, pp. 376–385. ACM (2018)
6. Du, P., Ogilvie, H., Nakhleh, L.: Unifying gene duplication, loss, and coalescence on phylogenetic networks. bioRxiv (2019). https://doi.org/10.1101/589655
7. Heled, J., Drummond, A.J.: Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. **27**(3), 570–580 (2009)

8. Jukes, T.H., Cantor, C.R., et al.: Evolution of protein molecules. Mamm. Protein Metab. **3**(21), 132 (1969)
9. Maddison, W.P.: Gene trees in species trees. Syst. Biol. **46**(3), 523–536 (1997)
10. Mallet, J.: Hybridization as an invasion of the genome. Trends Ecol. Evol. **20**(5), 229–237 (2005)
11. Mallet, J.: Hybrid speciation. Nature **446**(7133), 279 (2007)
12. Nakhleh, L.: A metric on the space of reduced phylogenetic networks. IEEE/ACM Trans. Comput. Biol. Bioinform. **7**(2), 218–222 (2010)
13. Nosil, P.: Speciation with gene flow could be common. Mol. Ecol. **17**(9), 2103–2106 (2008). https://doi.org/10.1111/j.1365-294X.2008.03715.x
14. Rambaut, A., Grass, N.C.: Seq-gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Bioinformatics **13**(3), 235–238 (1997)
15. Rasmussen, M., Kellis, M.: A Bayesian approach for fast and accurate gene tree reconstruction. Mol. Biol. Evol. **28**(1), 273–290 (2011)
16. Rasmussen, M.D., Kellis, M.: Unified modeling of gene duplication, loss, and coalescence using a locus tree. Genome Res. **22**(4), 755–765 (2012)
17. Robinson, D.F., Foulds, L.R.: Comparison of phylogenetic trees. Math. Biosci. **53**(1–2), 131–147 (1981)
18. Sjöstrand, J., Sennblad, B., Arvestad, L., Lagergren, J.: DLRS: gene tree evolution in light of a species tree. Bioinformatics **28**(22), 2994–2995 (2012)
19. Than, C., Ruths, D., Nakhleh, L.: PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinform. **9**(1), 1 (2008)
20. Wen, D., Nakhleh, L.: Co-estimating reticulate phylogenies and gene trees from multi-locus sequence data. Syst. Biol. **67**(3), 439–457 (2018)
21. Wen, D., Yu, Y., Hahn, M.W., Nakhleh, L.: Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. Mol. Ecol. **25**(11), 2361–2372 (2016). https://doi.org/10.1111/mec.13544
22. Wen, D., Yu, Y., Nakhleh, L.: Bayesian inference of reticulate phylogenies under the multispecies network coalescent. PLoS Genet. **12**(5), e1006006 (2016)
23. Wen, D., Yun, Y., Zhu, J., Nakhleh, L.: Inferring phylogenetic networks using PhyloNet. Syst. Biol. **67**(4), 735–740 (2018)
24. Zhang, B., Wu, Y.-C.: Coestimation of gene trees and reconciliations under a duplication-loss-coalescence model. In: Cai, Z., Daescu, O., Li, M. (eds.) ISBRA 2017. LNCS, vol. 10330, pp. 196–210. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59575-7_18
25. Zhang, C., Ogilvie, H.A., Drummond, A.J., Stadler, T.: Bayesian inference of species networks from multilocus sequence data. Mol. Biol. Evol. **35**(2), 504–517 (2018). https://doi.org/10.1093/molbev/msx307