

METHODOLOGY

Open Access

Integrated likelihood for phylogenomics under a no-common-mechanism model



Hunter Tidwell and Luay Nakhleh*

From 17th RECOMB Satellite Conference on Comparative Genomics
Montpellier, France. 1-4 October 2019

Abstract

Background: Multi-locus species phylogeny inference is based on models of sequence evolution on gene trees as well as models of gene tree evolution within the branches of species phylogenies. Almost all statistical methods for this inference task assume a common mechanism across all loci as captured by a single value of each branch length of the species phylogeny.

Results: In this paper, we pursue a “no common mechanism” (NCM) model, where every gene tree evolves according to its own parameters of the species phylogeny. Based on this model, we derive an analytically integrated likelihood of both species trees and networks given the gene trees of multiple loci under an NCM model. We demonstrate the performance of inference under this integrated likelihood on both simulated and biological data.

Conclusions: The model presented here will afford opportunities for exploring connections among various criteria for estimating species phylogenies from multiple, independent loci. Furthermore, further development of this model could potentially result in more efficient methods for searching the space of species phylogenies by focusing solely on the topology of the phylogeny.

Keywords: Phylogenomics, Multispecies coalescent, No common mechanism, Integrated likelihood

Background

A phylogenetic tree models the evolutionary history of a set of taxa (genes, species, etc.) from their most recent common ancestor. Analyses of genome-wide data from several groups of species have highlighted a significant phenomenon, namely the incongruence among phylogenetic trees of the different genomic regions as well as with the phylogeny of the species [1]. One cause of incongruence is incomplete lineage sorting, or ILS, which can be mathematically well understood under the *multispecies coalescent* (MSC) model [2]. Figure 1 illustrates the gene tree probability distribution that the multispecies coalescent defines.

However, as Maddison noted [1], other processes could give rise to incongruence among gene trees, including hybridization, which gives rise to phylogenetic networks [3]. The multispecies coalescent has been extended to incorporate such processes [4–7]. These findings have given rise to phylogenomics—the inference of a species phylogeny from genome-wide data. Given m gene trees $\mathcal{G} = \{g_1, \dots, g_m\}$ for m independent loci (genomic regions), the likelihood of a species phylogeny Ψ and its branch lengths Λ and inheritance probabilities Γ (more on these in the “Methods” section) is given by

$$\mathcal{L}(\Psi, \Lambda, \Gamma | \mathcal{G}) = \prod_{i=1}^m \mathbf{P}(g_i | \Psi, \Lambda, \Gamma). \quad (1)$$

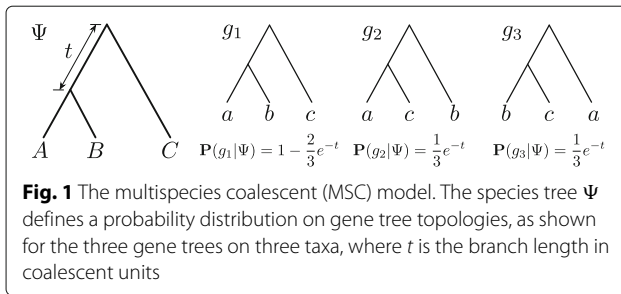
Assuming, for example, that ILS is the sole cause of all incongruence among gene trees in \mathcal{G} , then $\mathbf{P}(g_i | \Psi, \lambda)$ is

*Correspondence: nakhleh@rice.edu

¹Department of Computer Science, Rice University, Houston, TX, USA



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.



given by MSC, as illustrated in Fig. 1. If both ILS and reticulation are at play, then the probability distribution is given by the multispecies network coalescent [4].

The formulation given by Eq. (1) assumes a “common mechanism” across all loci—all gene trees “grow” within a species tree given by a single setting of branch lengths (and, in the case of a phylogenetic network, a single setting of inheritance probabilities). In this paper we explore a “no common mechanism,” or NCM, model, which states that each gene tree evolved within the branches of the species phylogeny under a totally separate process from all other gene trees as given by the parameters (e.g., branch lengths) of the species phylogeny. The main motivation behind this work is taming the complexity of statistical inference of phylogenetic networks. As calculating the likelihood of a phylogenetic network is computationally very expensive [8, 9], integrating out the continuous parameters and focusing the search on only the topologies could significantly reduce the computational requirements of phylogenetic network inference.

It is important to note that the NCM model has been explored and studied in the “classical” phylogeny problem (inferring a phylogenetic tree from a sequence alignment). Under that setting, NCM posits that each site in the sequence alignment evolved under its own branch lengths of the phylogenetic tree. Tuffley and Steel [10] established a seminal result in the field by proving that the maximum parsimony and maximum likelihood estimates of a phylogenetic tree are equal under an NCM model based on a symmetric Poisson process of nucleotide substitution. Additional mathematical results based on the NCM model were later established by Steel and Penny [11]. The NCM model allowed for analytically integrating out the branch lengths of a phylogenetic tree and efficiently exploring the space of phylogenetic trees [12]. Steel [13] showed that it is possible to achieve statistical consistency of inference under certain NCM models.

However, while the NCM model was used by some as an argument in favor of using maximum parsimony, Holder et al. [14] showed problems with this argument. Furthermore, Huelsenbeck et al. [15] argued that “biologically inspired phylogenetic models” outperform the NCM model. More specifically, it is hard to justify a phylogenetic model by which every site, including adjacent ones, has its

own evolutionary process. This is why the NCM model was deemed more useful for morphological characters than molecular characters in a sequence alignment. An NCM model is more appropriate in phylogenomics, where different genomic regions could have evolved under different rates and even under different trees. This has led to the development of methods that partition the genomic data into regions, where the sites within each region are assumed to have evolved identically; e.g., see [16]. In this paper, we analytically derive an integrated likelihood model under the multispecies coalescent with the NCM, and show the performance of inference based on this NCM model on data simulated under common and no common mechanisms, as well as a biological data set. Like the argument made by Huelsenbeck et al. [12], the work we present here could lead to more efficient ways of exploring the space of species phylogenies and establishing connections between inference based on different models, including the parsimony formulation given by the “minimizing deep coalescences” criterion [1, 17, 18]).

Methods

In order to account for both reticulation and incomplete lineage sorting, we use the phylogenetic network model since it generalizes trees. A *phylogenetic network* Ψ on set \mathcal{X} of taxa is a rooted, directed, acyclic graph (DAG) with set of nodes $V(\Psi) = \{r\} \cup V_L \cup V_T \cup R$, where

- $\text{indeg}(r) = 0$ (r is the root of Ψ);
- $\forall v \in V_L, \text{indeg}(v) = 1$ and $\text{outdeg}(v) = 0$ (V_L are the external tree nodes, or leaves, of Ψ);
- $\forall v \in V_T, \text{indeg}(v) = 1$ and $\text{outdeg}(v) \geq 2$ (V_T are the internal tree nodes of Ψ); and,
- $\forall v \in R, \text{indeg}(v) = 2$ and $\text{outdeg}(v) = 1$ (R are the reticulation nodes of Ψ).

The set of edges $E(\Psi) \subseteq V \times V$ consists of *reticulation edges*, whose heads are reticulation nodes, and *tree edges*, whose heads are tree nodes. The leaves of the network are bijectively labeled by elements of \mathcal{X} .

We assume that we have ℓ reticulation nodes $R = \{r_1, \dots, r_\ell\}$ with ℓ associated inheritance probabilities $\gamma_1, \dots, \gamma_\ell$, respectively (that is, node r_i has two parents pr_i and \overline{pr}_i with inheritance probabilities γ_i and $(1 - \gamma_i)$ associated with the edges $b1_i = (pr_i, r_i)$ and $b2_i = (\overline{pr}_i, r_i)$, respectively). In addition to the topology of a phylogenetic network Ψ , each edge $b = (u, v)$ in $E(\Psi)$ has a length λ_b measured in coalescent units, which is the number of generations divided by effective population size on that branch. We use Ψ to refer to the topology of the phylogenetic network, Λ to refer to its branch lengths, and Γ to refer to the inheritance probabilities associated with all reticulation nodes. A species tree is a phylogenetic network with no reticulation nodes (and an empty Γ).

Distribution of gene tree topologies

Given a phylogenetic network Ψ , its branch lengths Λ and inheritance probabilities Γ on the reticulation edges, the gene tree topology is a random variable whose probability mass function (pmf) we now briefly review. This pmf was originally derived for the case of species trees by Degnan and Salter [19] and later extended to the case of phylogenetic networks by Yu et al. [4].

We denote by Ψ_u the set of nodes that are reachable from the root of Ψ via at least one path that goes through node $u \in V(\Psi)$. Then given a phylogenetic network Ψ and a gene tree g for some locus j , a coalescent history is a function $h : V(g) \rightarrow E(\Psi)$ such that the following two conditions hold:

- if v is a leaf in g , then $h(v) = (x, y)$ where y is the leaf in Ψ with the label of the species from which the allele labeling leaf v in G is sampled; and,
- if v is a node in the subtree of g that is rooted at u , and $h(u) = (p, q)$, then $h(v) = (x, y)$ where $y \in \Psi_q$.

Given a phylogenetic network Ψ and a gene tree g for locus j , we denote by $H_\Psi(g)$ the set of all coalescent histories of g within the branches of Ψ . Then the pmf of the gene tree is given by

$$\mathbf{P}(g|\Psi, \Lambda, \Gamma) = \sum_{h \in H_\Psi(g)} \mathbf{P}(h|\Psi, \Lambda, \Gamma), \quad (2)$$

where Λ are the branch lengths of the phylogenetic network (in coalescent units), Γ is the inheritance probabilities matrix, and $\mathbf{P}(h|\Psi, \Lambda, \Gamma)$ gives the pmf of the coalescent history random variable, which can be computed as

$$\mathbf{P}(h|\Psi, \Lambda, \Gamma) = \frac{w(h)}{d(h)} \prod_{b \in E(\Psi)} \frac{w_b(h)}{d_b(h)} p_{u_b(h)v_b(h)}(\lambda_b) \prod_{r_i \in R} \gamma_i^{u_{b1_i}(h)} (1 - \gamma_i)^{u_{b2_i}(h)}. \quad (3)$$

In this equation, $u_b(h)$ and $v_b(h)$ denote the number of lineages that respectively enter and exit edge b of Ψ under coalescent history h . The term $p_{u_b(h)v_b(h)}(\lambda_b)$ is the probability of $u_b(h)$ gene lineages coalescing into $v_b(h)$ during time λ_b . And $w_b(h)/d_b(h)$ is the proportion of all coalescent scenarios resulting from $u_b(h) - v_b(h)$ coalescent events that agree with the topology of the gene tree. This quantity without the b subscript corresponds to the root of Ψ . Notice that removing the rightmost product over the reticulation nodes in Eq. (3) gives the pmf for species trees.

An integrated likelihood framework

Integrating out the branch lengths. The function $p_{uv}(t)$ employed by Eq. (3) is given by

$$p_{uv}(t) = \sum_{j=u}^v \left(e^{-\frac{j(j-1)}{2}t} \frac{(2j-1)(-1)^{j-v}}{v!(j-v)!(v+j-1)} \prod_{y=0}^j \frac{(v+y)(u-y)}{u+y} \right), \quad (4)$$

which, for simplifying the equations below, can be written as $p_{uv}(t) = \sum_{j=u}^v \left(e^{-\frac{j(j-1)}{2}t} \cdot f(u, v, j) \right)$, where $f(u, v, j) = \frac{(2j-1)(-1)^{j-v}}{v!(j-v)!(v+j-1)} \prod_{y=0}^j \frac{(v+y)(u-y)}{u+y}$. Assuming a truncated Exponential prior with support in $(0, \tau]$ and hyperparameter value of 1 on t , we have

$$\begin{aligned} p_{uv}^\tau &= \int_0^\tau p_{uv}(t) p(t) dt \\ &= \int_0^\tau \left(\sum_{j=u}^v \left(e^{-\frac{j(j-1)}{2}t} \cdot f(u, v, j) \right) \right) \left(\frac{e^{-t}}{1-e^{-\tau}} \right) dt \\ &= \frac{1}{1-e^{-\tau}} \sum_{j=u}^v f(u, v, j) \int_0^\tau e^{-\left(\frac{j(j-1)}{2}+1\right)t} dt \\ &= \frac{1}{1-e^{-\tau}} \sum_{j=u}^v \left(f(u, v, j) \cdot \frac{2}{j(j-1)+2} \left(1 - e^{-\left(\frac{j(j-1)}{2}+1\right)\tau} \right) \right). \end{aligned} \quad (5)$$

Using this result, and assuming all branch lengths of the species phylogeny are independent, we have

$$\begin{aligned} \mathbf{P}(h|\Psi, \Gamma) &= \int_0^\tau \mathbf{P}(h|\Psi, \Lambda, \Gamma) p(\Lambda) d\Lambda \\ &= \frac{w(h)}{d(h)} \prod_{b \in E(\Psi)} \frac{w_b(h)}{d_b(h)} p_{u_b(h)v_b(h)}^\tau \prod_{r_i \in R} \gamma_i^{u_{b1_i}(h)} (1 - \gamma_i)^{u_{b2_i}(h)}. \end{aligned} \quad (6)$$

Integrating out the inheritance probabilities. We now have $\mathbf{P}(h|\Psi) = \int \mathbf{P}(h|\Psi, \Gamma) p(\Gamma) d\Gamma$, where $p(\Gamma)$ is a prior on the inheritance probabilities, and the multiple integration is taken over all ℓ gamma's on $[0, 1]$. We assume $\gamma_i \sim \text{Beta}(2, 2)$, so that we have a conjugate prior (pdf in this case is $\frac{\gamma_i(1-\gamma_i)}{B(2,2)}$). Then, we have

$$\begin{aligned} \mathbf{P}(h|\Psi) &= \frac{w(h)}{d(h)} \prod_{b \in E(\Psi)} \frac{w_b(h)}{d_b(h)} p_{u_b(h)v_b(h)}^\tau \\ &\quad \prod_{r_i \in R} \frac{1}{B(2,2)} \int_0^1 \gamma_i^{u_{b1_i}(h)+1} (1 - \gamma_i)^{u_{b2_i}(h)+1} d\gamma_i \\ &= \frac{w(h)}{d(h)} \left(\prod_{b \in E(\Psi)} \frac{w_b(h)}{d_b(h)} p_{u_b(h)v_b(h)}^\tau \right) \left(\frac{1}{B(2,2)} \right)^\ell \\ &\quad \prod_{r_i \in R} \frac{(u_{b1_i}(h)+1)!(u_{b2_i}(h)+1)!}{(u_{b1_i}(h)+u_{b2_i}(h)+3)!}. \end{aligned} \quad (7)$$

Finally,

$$\mathbf{P}(g|\Psi) = \sum_{h \in H_\Psi(g)} \mathbf{P}(h|\Psi), \quad (8)$$

where Ψ is given by its topology alone.

Observe that if one treats the branch lengths and inheritance probabilities in Eq. (1) as a nuisance parameter, then the integrated likelihood is given by $\mathbf{P}(\mathcal{G}|\Psi) = \int \int \left[\prod_{i=1}^m \mathbf{P}(g_i|\Psi, \Lambda, \Gamma) p(\Lambda) p(\Gamma) d\Lambda d\Gamma \right]$, where $\mathcal{G} = \{g_1, g_2, \dots, g_m\}$, and $\mathbf{P}(g_i|\Psi, \lambda)$ is computed as in [4, 20].

Inference The calculation given by Eq. (8) above allows us to compute

$$f(\Psi|\mathcal{G}) = \prod_{i=1}^m \left[\int_{\Lambda} \int_{\Gamma} \mathbf{P}(g_i|\Psi, \Lambda, \Gamma) p(\Lambda) p(\Gamma) d\Lambda d\Gamma \right] = \prod_{i=1}^m \mathbf{P}(g_i|\Psi) \quad (9)$$

from a set \mathcal{G} of input gene trees inferred on multiple independent loci. For inferring an optimal network under Eq. (9), also known as the maximum integrated likelihood network, a search for the network Ψ that maximizes $f(\Psi|\mathcal{G})$ is conducted. Since no branch lengths or inheritance probabilities are optimized or sampled, just like the case of inference under the MDC criterion, we use the exact search heuristic and moves of [18].

Efficiency Observe that the computational requirements of calculating the probability according to Eq. (8) with the analytical integration of branch lengths and inheritance probabilities remain the same as those of computing the probability of a gene tree given a species network and its specific branch lengths and inheritance probabilities as in [4]. The major gain in computational requirements is in the search procedure. Search based on the integrated likelihood evaluates only topologies, and need not consider optimizing or sampling the continuous parameters. For example, for the case of a rooted species tree on three taxa, search based on the integrated likelihood only inspects three topologies; that is, likelihood calculations are done exactly three times to identify the optimal species tree. Searching for the maximum likelihood species tree while sampling branch lengths requires walk in the infinite space of branch length settings. It is important to note that the likelihood of different parameterizations of

a given species tree (or network) is not a “nice” convex function. Therefore, searching for branch lengths and, in the case of phylogenetic networks, inheritance probabilities that maximize the likelihood is not a simple computational task and requires dealing with local maxima. Using integrated likelihood, it is even possible to evaluate phylogenetic networks with small numbers of taxa even exhaustively, a task that cannot be done once the branch lengths and inheritance probabilities are involved. For example, there are 105 rooted species trees on five taxa. Finding the one that maximizes the integrated likelihood based on the computations above can be done by exhaustively calculating the likelihoods for all 105 tree topologies. When branch lengths are included, searching the space while optimizing or sampling the branch lengths cannot be avoided.

Results

Accuracy on data simulated under a common mechanism

We first set out to study the performance of maximum integrated likelihood inference under the NCM model, and compare it to that of inference under the parsimony criterion “minimizing deep coalescence” (MDC) of [18]. We follow the same simulation setup, including the model networks, parameters, and numbers of gene trees as that in [18]. More specifically, we considered four phylogenetic network topologies involving distinct combinations of reticulation and speciation events, as shown in Fig. 2. To better understand the effects of deep coalescence in each scenario, we used two settings of the branch length parameters for each network. In branch length setting 1, each of the values t_1 , t_2 , t_3 , and t_4 are equal to 1 coalescent unit. In branch length setting 2, each value is equal to 2 coalescent units. Setting 1 should involve

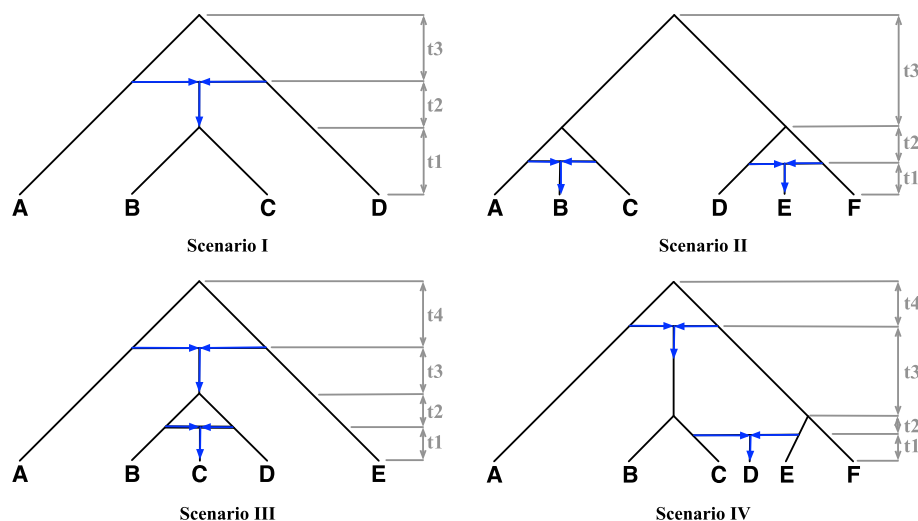


Fig. 2 Model phylogenetic networks. Blue arrows indicated directions into and out of the reticulation nodes

more deep coalescence events, while setting 2 involves longer branches which are less likely according to the exponential prior on the branch lengths. Each provides unique challenges for the integrated likelihood inference under NCM. All inheritance probabilities were set to 0.5. For each setting and for each number of loci in the set {10, 25, 50, 100, 500, 1000, 2000}, we generated 100 data sets of gene trees using the program *ms* [21]. It is important to highlight here that the data was generated under a common mechanism; that is, not under the NCM model. We then ran the inference method under MDC of [18] and the maximum integrated likelihood inference under the NCM model on each data set. Since neither the MDC nor the likelihood criteria allow for determining model complexity (in this case, the number of reticulations) in a systematic way, we ran the method here with the maximum number of reticulations set at the true number, which is 1 for the data corresponding to Scenario I, and 2 for the other three scenarios. We then computed the topological distance of [22], as implemented in *PhyloNet* [23], between each inferred network and the model network on which the data was generated, and averaged the results over all 100 data sets for each setting. The results are shown in Fig. 3. Note that for the calculation of the NCM integrated likelihood, we used a non-truncated exponential prior for the branch lengths of the species phylogeny. This is equivalent to letting the

hyperparameter τ grow arbitrarily large, and results in a similar likelihood function as that in Eq. (5).

As the results show, both methods have almost the same behavior and accuracy under branch length setting 1 for the networks of Scenarios I, III, and IV, and under branch length setting 2 for the network of Scenario IV. However, while MDC always converged onto the true network, inference under the NCM model diverged from the true network in the other cases. We then set out to compare the erroneous networks inferred under the NCM model (Fig. 4) to their true counterparts. A quick inspection of the three networks in Fig. 4 points to a very interesting pattern. The only errors in the inferences were the direction of the reticulation edge (as highlighted with red arrows in the figure). When we inspect the three networks in Fig. 4 and their counterparts in Fig. 2, we find that every two corresponding networks in the two figures display the same set of trees. Each of the network corresponding to Scenario I in Fig. 2 and the network in Fig. 4a displays the same two trees: $((A,(B,C)),D)$ and $(A,((B,C),D))$. Each of the network corresponding to Scenario II in Fig. 2 and the network in Fig. 4b displays the same four trees $((A,B),C),((D,E),F)), (((A,B),C),(D,(E,F))), ((A,(B,C)),((D,E),F))$, and $((A,B),C),(D,(E,F)))$. Each of the network corresponding to Scenario III in Fig. 2 and the network in Fig. 4c displays the same four trees: $((A,((B,C),D)),E)$, $((A,(B,(C,D))),E)$, $((E,((B,C),D)),A)$, and

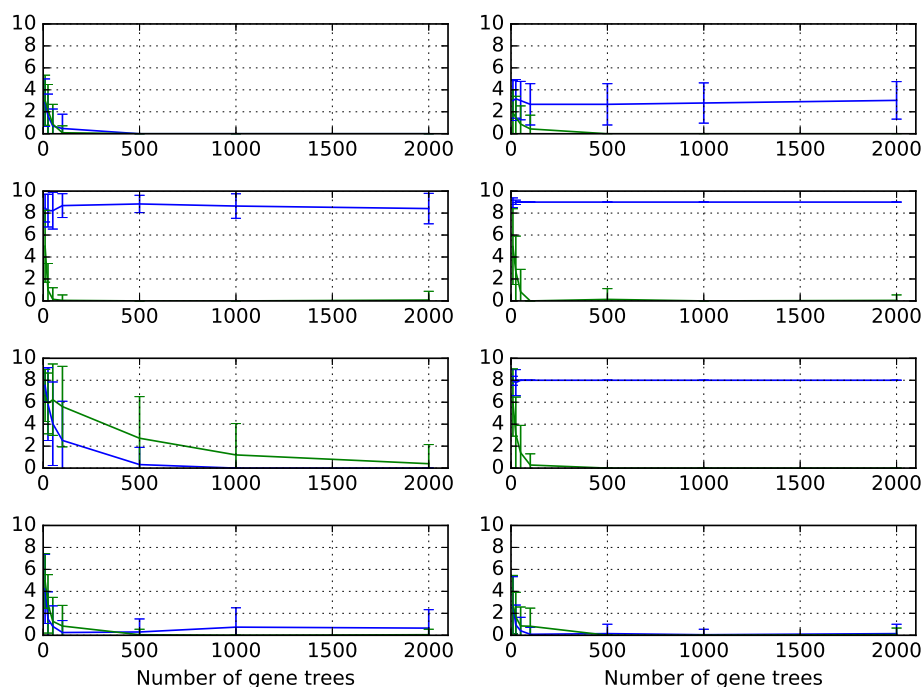


Fig. 3 Accuracy of network inference on data simulated under a common mechanism. The symmetric network difference between the inferred and model network, averaged over 100 trials, using the MDC criterion as implemented in [18] (green) and the maximum integrated likelihood under the NCM model (blue). Rows from top to bottom correspond to Scenarios I-IV, respectively, of Fig. 2. Left and right columns correspond to branch length settings 1 and 2, respectively

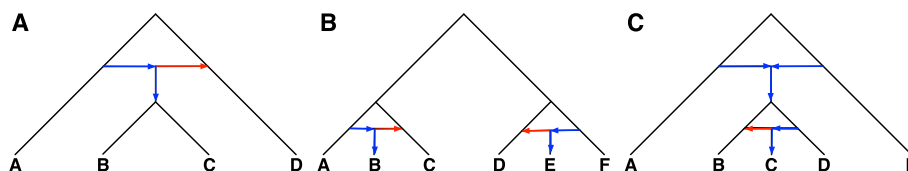


Fig. 4 The incorrect networks inferred under the NCM model. While the correct networks were inferred for many data sets, incorrect networks were inferred in other cases, and those incorrect networks are shown in this figure. **a** The network inferred from the data generated on the network of Scenario I and branch length setting 2. **b** The network inferred from the data generated on the network of Scenario II and both branch length settings. **c** The network inferred from the data generated on the network of Scenario III and branch length setting 2. Red arrows indicate the reticulations whose direction was inferred in the reverse order

((E,(B,(C,D))),A). These pairs of networks are indistinguishable when a single individual per species is sampled, as discussed in [24]. Branch length information on the gene trees and/or sampling multiple individuals per species could resolve this indistinguishability [25], especially as the set of displayed trees does not necessarily characterize a phylogenetic network in the presence of ILS [8].

Accuracy on data simulated without a common mechanism

To study the behavior of inference under an NCM model when the data are evolved without a common mechanism, we again used the network topologies in Fig. 2 to simulate the data. To model a no-common-mechanism evolutionary process, in this experiment, each time a single gene tree was simulated under a network topology, settings for the continuous parameters were sampled from a distribution. As before, to vary the amount of deep coalescence, we used two settings for the branch lengths of each network topology. In branch length setting 1, each of the values t_1 , t_2 , t_3 , and t_4 are sampled from a uniform distribution from 0 to 2 coalescent units. In branch length setting 2, each value is sampled from a uniform distribution from 0 to 4 coalescent units. In each case, every inheritance probability was sampled from a uniform distribution from 0 to 1. For each setting and for each number of loci in the set {10, 25, 50, 100, 500, 1000, 2000}, we generated 20 data sets of gene trees. We again inferred a network for each collection of data using both MDC and maximum integrated likelihood under NCM, and calculated each average topological difference to the true network. The results are shown in Fig. 5. As the results show, except for Scenario II, inference under the NCM model now improved, and for branch length setting 1 on Scenario III, MDC's performance became poor.

Analysis of a mosquito data set

We reanalyzed the *Anopheles* data of [26]. The data consist of one genome from each of the species *An. gambiae* (gam), *An. coluzzii* (col), *An. arabiensis* (ara), *An. quadriannulatus* (qua), *An. merus* (mer) and *An. melas* (mel). *An. christyi* serves as the outgroup for rooting the

gene trees. We used the same set of gene trees from the autosomes that were used in the analyses of [27]. In particular, we used the same set of 669, 849, 564, and 709 loci from the 2L, 2R, 3L, and 3R chromosomes, respectively, and where 100 maximum likelihood bootstrap trees were inferred for each locus and used in the inference. These data are already available in DRYAD, entry doi:10.5061/dryad.tn47c. In [27], the phylogenetic network was inferred from the gene trees using the maximum likelihood method of [5]. The phylogenetic networks from the original study of [26], from the maximum likelihood analysis of [27], and the one we obtained under the NCM model are shown in Fig. 6.

As reported in [27], the likelihood of the network in Fig. 6b was much higher than that of the network reported by Fontaine et al. and shown in Fig. 6a. The log-likelihood under the NCM model of the networks in Fig. 6b and 6c are -17520.49 and -16821.33, respectively. This demonstrates that the difference in the inferred networks is not due to limitations of the search procedure, but due to a better likelihood of the new inferred network over existing ones under the NCM model. It is important to note both networks of Fig. 6b and 6c agree on the same underlying tree structure (the one obtained from the network by removing the green horizontal arrows, and sometimes called the backbone tree). This tree disagrees with that reported by Fontaine et al. [26], and this disagreement was discussed in [27]. Our inferred network also agrees with that of [27] in terms of the *An. quadriannulatus* and *An. merus* hybridization. However, the *An. merus* and *An. melas* hybridization differ in terms of the direction of the reticulation (similar to the trend observed on the simulated data and discussed above), and the *An. quadriannulatus* and *An. melas* hybridization is not reported in [27].

There could be several reasons for the differences between the two networks of Fig. 6b and 6c. The obvious one is that the two networks were inferred under two different models, one that a common mechanism of evolution underlies all loci and the other that assumes each locus has its own model. Second, as discussed in [28], (unpenalized) maximum likelihood cannot determine the

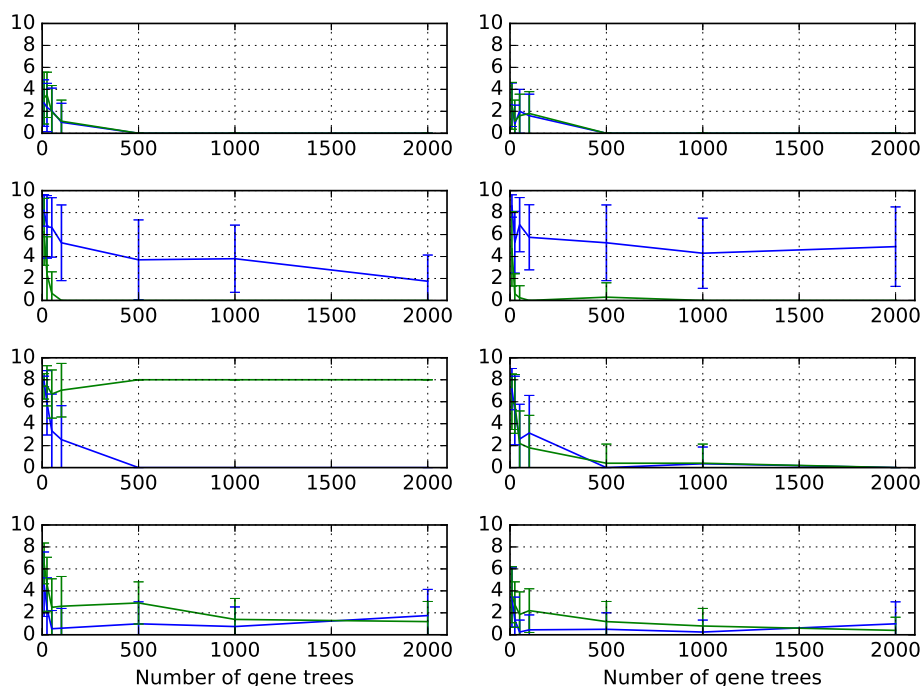


Fig. 5 Accuracy of network inference on data simulated under NCM. The symmetric network difference between the inferred and model network, averaged over 20 trials, using the MDC criterion as implemented in [18] (green) and the maximum integrated likelihood under the NCM model (blue). Rows from top to bottom correspond to Scenarios I-IV, respectively, of Fig. 2. Left and right columns correspond to branch length settings 1 and 2, respectively

correct number of reticulations. It could be that as more complex networks (ones with more than three reticulations) are inferred, the two analyses using the method of [5] and the one under the NCM model might converge onto the same network. These differences notwithstanding, inference under the NCM model recovered a very similar network, which makes it promising to explore the network space first without optimizing or sampling the continuous parameters, and then potentially follow up with a sampling phase to recover parameter values.

MDC vs. integrated likelihood under nCM

Tuffley and Steel [10] showed that an NCM model is related to Fitch's parsimony for character evolution in the following way. The maximum likelihood tree (or trees)

under their NCM model is also a maximum parsimony tree under certain conditions. In light of this, we investigated whether a similar correspondence might hold for the NCM model for gene trees evolving down a species trees. Using the three gene trees of Fig. 7a, and assuming equal frequencies of all three, we inferred the optimal tree under the MDC criterion as well as the optimal tree under the NCM model. Figure 7b-d shows the results. As the results show, the two criteria in this case result in different optimal species trees. Furthermore, in this case, the optimal tree under the NCM model is not unique.

Conclusions

In this paper, we introduced a no-common mechanism for phylogenomics, where the species phylogeny topology is the same across all loci, but the gene tree of every

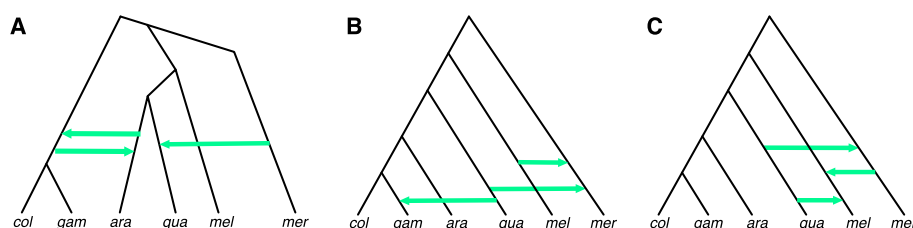


Fig. 6 Networks for the mosquito data set. **a** The phylogenetic network reported in [26]. **b** The phylogenetic network analyzed using the maximum likelihood method of [5] and reported in [27]. **c** The phylogenetic network inferred under the NCM model

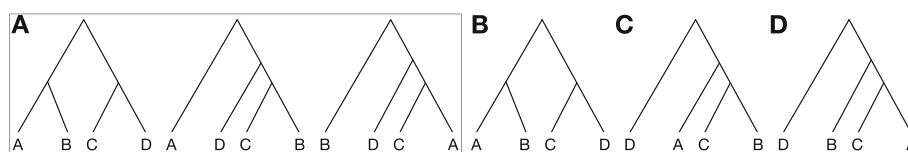


Fig. 7 Different optimal species trees under the MDC criterion and NCM model. **a** Three gene trees assumed to have equal frequencies. **b** Optimal species tree under the MDC criterion (has a cost of 4 extra lineages). The other two trees each have a cost of 5 extra lineages. **c** and **d** Optimal species trees under the NCM model

locus evolves under its own parameter (branch lengths and inheritance probabilities) settings. We implemented a maximum integrated likelihood function under this NCM model, assessed its accuracy and compared it to inferences under the parsimony MDC criterion on simulated data and the maximum likelihood inference on an empirical data set of mosquito genomes. We found that the inference produces very good results and when there is a disagreement, it is most often due to indistinguishability when using only gene tree topologies and a single individual per species. The main rationale behind developing such a mechanism is to allow for developing methods for efficiently exploring the species phylogeny space while focusing on traversing different topologies without the need for sampling or optimizing branch lengths and other continuous parameters. The running time of calculating the likelihoods (integrated and standard) of the networks in the simulated data above took few than 10 minutes for each network (and the set of gene trees used as the input data). The major gain is in the running time of maximum likelihood inference based on the two methods. The standard maximum likelihood estimation was run for eight hours on each data set, and it took all that time to evaluate candidate networks with their branch lengths. In the case of inference based on the integrated likelihood, and since branch lengths do not factor in the search itself, it took much less time to infer (locally) optimal networks. In fact, if a search technique is designed so as to ensure that the same network topology is not visited more than once during the search (which is not the case in the heuristic employed currently in PhyloNet), search based on the integrated likelihood would be improved much more significantly.

The work gives rise to several questions. First, under which conditions are optimal trees or networks identical under both the MDC criterion and the NCM model? Second, is there a parsimony criterion other than MDC under which the optimal species phylogeny is always identical to the optimal tree under the NCM model? Third, under what conditions, if any, is inference under the NCM model statistically consistent? Fourth, in light of the results above, an important question would be to explore why optimal networks under the NCM model often have

reticulations in the opposite direction from those in optimal networks under the likelihood model with a common mechanism across all loci. Answering these and other questions will open up many research avenues in the area of phylogenomics.

Finally, it is important to conduct more studies where different loci have different evolutionary parameters so as to mimic data coming from autosomes and sex chromosomes, as well as loci under selection or where duplication and loss could have played a role. In such cases, assuming a common mechanism underlying all loci is inappropriate and inference under an NCM model could provide more accurate results.

Abbreviations

DAG: Directed, acyclic graph; ILS: Incomplete lineage sorting; MDC: Minimizing deep coalescence; MSC: Multispecies coalescent; NCM: No common mechanism

Acknowledgements

The authors thank Jiafan Zhu for help with implementing the method and running the experiments, and three anonymous reviewers for detailed comments that helped improve the manuscript.

About this supplement

This article has been published as part of BMC Genomics Volume 21 Supplement 2, 2020: Proceedings of the 17th 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-21-supplement-2>.

Authors' contributions

Both authors designed the study. HT implemented the method and produced the results. Both authors analyzed the results and wrote the manuscript. Both read and approved the final manuscript.

Funding

This work was supported in part by grants DBI-1355998, CCF-1302179, CCF-1514177, CCF-1800723, and DMS-1547433 from the National Science Foundation of the United States of America. Publication costs for this work are also funded by these grants.

Availability of data and materials

Not applicable.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Published: 16 April 2020

References

1. Maddison W. Gene trees in species trees. *Syst Biol.* 1997;46(3):523–36.
2. Degnan JH, Rosenberg NA. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol Evol.* 2009;24(6):332–40.
3. Nakhleh L. Evolutionary phylogenetic networks: models and issues. In: *Problem Solving Handbook in Computational Biology and Bioinformatics*. Springer; 2010. p. 125–158. https://doi.org/10.1007/978-0-387-09760-2_7.
4. 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.
5. Yu Y, Dong J, Liu K, Nakhleh L. Maximum likelihood inference of reticulate evolutionary histories. *Proc Natl Acad Sci.* 2014;111(46):16448–53.
6. Wen D, Yu Y, Nakhleh L. Bayesian inference of reticulate phylogenies under the multispecies network coalescent. *PLoS Genet.* 2016;12(5):1006006.
7. Wen D, Nakhleh L. Co-estimating reticulate phylogenies and gene trees from multi-locus sequence data. *Syst Biol.* 2018;67(3):439–57.
8. Zhu J, Yu Y, Nakhleh L. In the light of deep coalescence: revisiting trees within networks. *BMC Bioinformatics.* 2016;17(14):415.
9. Elworth RAL, Ogilvie HA, Zhu J, Nakhleh L. Bioinformatics and Phylogenetics: Seminal Contributions of Bernard Moret. In: Warnow T., editor. Cham: Springer; 2019. p. 317–360.
10. Tuffley C, Steel M. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bull Math Biol.* 1997;59(3):581–607.
11. Steel M, Penny D. Parsimony, likelihood, and the role of models in molecular phylogenetics. *Mol Biol Evol.* 2000;17(6):839–50.
12. Huelsenbeck JP, Ane C, Larget B, Ronquist F. A Bayesian perspective on a non-parsimonious parsimony model. *Syst Biol.* 2008;57(3):406–19.
13. Steel M. Can we avoid SIN in the house of no common mechanism?. *Syst Biol.* 2010;60(1):96–109.
14. Holder MT, Lewis PO, Swofford DL. The Akaike information criterion will not choose the no common mechanism model. *Syst Biol.* 2010;59(4):477–85.
15. Huelsenbeck JP, Alfaro ME, Suchard MA. Biologically inspired phylogenetic models strongly outperform the no common mechanism model. *Syst Biol.* 2011;60(2):225–32.
16. Angelis K, Álvarez-Carretero S, Dos Reis M, Yang Z. An evaluation of different partitioning strategies for bayesian estimation of species divergence times. *Syst Biol.* 2017;67(1):61–77.
17. Than C, Nakhleh L. Species tree inference by minimizing deep coalescences. *PLoS Comput Biol.* 2009;5(9):1000501.
18. Yu Y, Barnett RM, Nakhleh L. Parsimonious inference of hybridization in the presence of incomplete lineage sorting. *Syst Biol.* 2013;62(5):738–51.
19. Degnan JH, Salter LA. Gene tree distributions under the coalescent process. *Evolution.* 2005;59:24–37.
20. Yu Y, Ristic N, Nakhleh L. Fast algorithms and heuristics for phylogenomics under ILS and hybridization. *BMC Bioinformatics.* 2013;14(Suppl 15):6.
21. Hudson RR. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics.* 2002;18:337–38.
22. Nakhleh L. A metric on the space of reduced phylogenetic networks. *IEEE/ACM Trans Comput Biol Bioinforma (TCBB).* 2010;7(2):218–222.
23. Than C, Ruths D, Nakhleh L. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics.* 2008;9(1):322.
24. Pardi F, Scornavacca C. Reconstructible phylogenetic networks: do not distinguish the indistinguishable. *PLoS Comput Biol.* 2015;11(4):1004135.
25. Zhu S, Degnan JH. Displayed trees do not determine distinguishability under the network multispecies coalescent. *Syst Biol.* 2016;66(2):283–98.
26. Fontaine MC, Pease JB, Steele A, Waterhouse RM, Neafsey DE, Sharakhov IV, Jiang X, Hall AB, Catteruccia F, Kakani E, Mitchell SN, Wu

- Y-C, Smith HA, Love RR, Lawniczak MK, Slotman MA, Emrich SJ, Hahn MW, Besansky NJ. Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science.* 2015;347(6217):1258524.
27. 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;25:2361–72.
28. Wen D, Yu Y, Zhu J, Nakhleh L. Inferring phylogenetic networks using PhyloNet. *Syst Biol.* 2018;67(4):735–40.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

