Referee report on Host-symbiont-gene phylogenetic reconciliation, by Menet et al. (https://doi.org/10.1101/2022.07.01.498457)

This manuscript proposes a 3-level probabilistic co-evolution model, to reconciliate the phylogenies of host species with their symbionts species and the genes of the latter. The model generalizes the method called ALE (Amalgamated likelihood estimation) in its version for undated trees, that enables exploring reconciliations between 2 trees under a Duplication-Transfer-Loss model of co-evolution. When considering 3-level undated phylogenies, the new model proposed by the authors enables considering duplications, transfers and losses of the symbionts within their host species (with probabilities θ_H that are fixed and not estimated by the model; rather these are pre-estimated through an Expectation-Maximization algorithm), together with duplications, intra-transfer and losses of the genes inside the symbiont species (probabilities θ_S). Here, intra-transfer means that a gene may transfer only between 2 symbionts that are within the same host species at the time of the transfer. Nonetheless, as in ALE, the method includes the possibility to use so-called ghost lineages for (indirect) transfers of genes between symbionts not present in the same host. Note that the method does not check for time feasibility, so it can output unvalid reconciliations.

The authors develop an algorithm for approximating the likelihood of any dataset (trees of hosts, symbionts and their genes) and inferring the parameters of the gene/symbiont co-evolution, relying on two versions of their method (Monte Carlo approximation with samples of reconciliations from the symbiont tree to the host tree; or sequential approach that relies on the most likely reconciliation from the symbiont tree to the host tree). When the symbiont tree is unknown, they also propose an option to infer it by amalgamation. In practice, the method is applied with many gene families (thus many gene trees).

Simulations under an external model are proposed, and the authors compare the 2 versions of their method (sequential and Monte-Carlo based) with a 2-level reconciliation of genes tree in their symbionts tree. Performance is measured with respect to the capacity of the 3 methods to recover gene transfers between correct symbiont donor and symbiont recipient (precision and recall are weighted wrt estimated probability of each transfer). The difference between the likelihoods of symbiont/gene reconciliations in the 3-level approach and in the 2-level one is used as a measure of host/symbiont co-evolution. Finally, the method is illustrated on 2 datasets: a Cinara aphids enterobacteria system and Helicobacter pylori within humans.

This is an important contribution to the 3-level reconciliation problem. The remarks below should help the authors clarify some points.

Major remarks

1. There is a confusion in the text between most likely reconciliation and maximum likelihood. I detail the problematic points below.

- When describing the 2-level reconciliation model (line 137 and below), the authors write: "*We do not have to enumerate all scenarios to compute that sum, because we can compute this likelihood using dynamic programming, considering matching all couples of gene and species sub-trees, starting from the leaves, and enumerating all possible events to get each match.*" This is not correct. Dynamic programming is a way to compute, for any parameter value $\theta_S = (p_S^S, p_S^D, p_S^T, p_S^L)$ the quantity

$$\max_{r_{G,S} \in R_{G,S}} \mathbb{P}_{\theta_S}(G, S, r_{G,S}), \tag{1}$$

but this quantity is different from the model likelihood, that equals the sum over all possible reconciliations

$$\mathbb{P}_{\theta_S}(G,S) = \sum_{r_{G,S}\in R_{G,S}} \mathbb{P}_{\theta_S}(G,S,r_{G,S}). \tag{2}$$

Dynamic programming algorithm constructs a table of all possible successive events from the leaves to the root, together with pointers that indicate at each stage the most likely event (for a fixed parameter value θ_S). To obtain the exact likelihood of the data, one should enumerate all possible paths (i.e. reconciliations) within that table and sum the corresponding probabilities; while backtracking in this table only outputs the most likely path (i.e. reconciliation). So if I understood correctly, at this stage (of the reconciliation between G and S) rather than *sampling* reconciliation scenarios, the authors *compute the most likely one*, say $\hat{r}_{G,S}$ that realizes the maximum in Eq (1) (for any parameter value θ_S), thanks to dynamic programming. While the chosen strategy makes sense, it's nonetheless different from a maximum likelihood one, where one would estimate θ_S by considering the argmax over θ_S of Eq (2).

- I believe that one layer of reconciliation is missing in the equations presented in Section 3.2. As far as I understand, Eq.(1) should be modified in the following way (I also added as indexes of the probabilities the different parameters θ_S and θ_H, for more clarity)

$$\begin{aligned} P_{(\theta_S,\theta_H)}(G|S,H) &= \sum_{r_{S,H}\in R_{S,H}} P_{\theta_S}(G|S,H,r_{S,H}) P_{\theta_H}(r_{S,H}|S,H) \\ &= \sum_{r_{S,H}\in R_{S,H}} \sum_{r_{G,S}\in R_{G,S}} P_{\theta_S}(G,r_{G,S}|S,H,r_{S,H}) P_{\theta_H}(r_{S,H}|S,H), \\ &\simeq \sum_{r_{S,H}\in R_{S,H}} P_{\theta_S}(G,\hat{r}_{G,S}|S,H,r_{S,H}) P_{\theta_H}(r_{S,H}|S,H), \end{aligned}$$

where $\hat{r}_{G,S}$ is the most likely reconciliation of G in S (for the current parameter value θ_S and the fixed reconciliation $r_{S,H}$). This first approximation makes sense since the most likely reconciliation $\hat{r}_{G,S}$ contributes to the dominant term in the sum $\sum_{r_{G,S}\in R_{G,S}}$ and one hopes the other terms are negligible. Then, if I understand correctly, a sequence $r_n \in R_{S,H}$ of reconciliations of the symbiont tree S within the host tree H is sampled and the authors make the second approximation of the likelihood through

$$P_{(\theta_S,\theta_H)}(G|S,H) \simeq \frac{1}{N}\sum_{n=1}^{N} P_{\theta_S}(G,\hat{r}_{G,S}|S,H,r_n) P_{\theta_H}(r_n|S,H). \tag{3}$$

In any case, Eq. (2) in the manuscript is not correct and a weight $P_{\theta_H}(r_n|S,H)$ is missing in that equation. To summarize, I understood that (in the first version of the algorithm, the sequential one being different) the authors sample a reconciliation $r_n \in R_{S,H}$; compute its probability $P_{\theta_H}(r_n|S,H)$ (thanks to a dynamic programming table); then they find the most likely reconciliation $\hat{r}_{G,S} \in R_{G,S}$ (thanks to a second dynamic programming table) that maximizes the probability $P_{\theta_S}(G,\hat{r}_{G,S}|S,H,r_n)$, together with the corresponding maximum value of that probability. (Note that this most likely reconciliation depends on the parameter θ_S and on the sampled reconciliation r_n). By doing this for many sampled reconciliations r_n, the authors finally compute the approximation in the right-hand side of Eq. (3). This quantity may be computed for a fixed parameter value (θ_S,θ_H) and the authors search for its maximum wrt (θ_S,θ_H). (In fact, they will pre-estimate θ_H with the Expectation-Maximization approach implemented in ALE; and then output mean a posteriori values for θ_S by sampling reconciliations of the gene/symbiont trees).

- In the sequential version of their method (Section 3.4), I understand that the authors now consider the following approximation

$$P_{(\theta_S,\theta_H)}(G|S,H) \simeq P_{\theta_S}(G,\hat{r}_{G,S}|S,H,\hat{r}_{S,H})P_{\theta_H}(\hat{r}_{S,H}|S,H),$$

  where $\hat{r}_{S,H}$ is the most likely reconciliation between the symbiont tree and the host tree. If this is indeed the case, it could be useful to write it down.

2. The authors choose not to produce simulations under their own model (Line 165). While it's interesting to use an external model as they did, that does not replace the simulations under the true model, to evaluate both the estimation procedure and potential identification issues in the model. Indeed, as the reconciliation models become more and more elaborate, the issue of knowing what portion of information about the past co-evolutionary events remains as a signal in the data is crucial. This can only be assessed through scenarios under the model at stake.

Minor remarks

- Line 278: "*In consequence we cannot use the efficient computation trick used for uniform rates.*" Please give a reference for that trick.
- Line 287: and below: should be made explicit that the times are given for the sequential version.
- Line 323: "*We did that by adding the symbiont tree as a possible host tree*". That sentence suggests that many host trees can be input in the method. However, I think this has not been said before. Please clarify this point.

Typos

- Line 218, $P(e \to h)$ should be $P^T(e \to h)$.
- Line 413, "*the 1.0 model* what does that mean?