Detecting outliers in heterochronous phylogenetic trees

Richard Neher

(Dated: January 3, 2025)

Misdating of sequences, or sequences with many sequencing errors, commonly distort time scaled phylogenetic trees. A common tactic to spot such sequences is to plot the root-to-tip distance as a function of time, reroot to optimize the correlation between them, and exclude tips that are far from the regression line. This is what augur, treetime, and TempEst currently do.

The problem with this approach is that it is not very sensitive since it ignores phylogenetic relationships among the tips. A sequence dated a year too early might still fall into the distribution of root-to-tip distances of that date, but is a clear outlier when compared directly to its neighbors in the tree.

To spot such outliers sensitively, we model the distribution in time for samples of a particular genotype i as

$$P(t|\tau_i, \sigma) = e^{-\frac{(t-\tau_i)^2}{2\sigma^2}} / \sqrt{2\pi\sigma^2} \tag{1}$$

Here τ_i is the time when most samples of this genotype are around, σ is the width of this distribution, which could correspond to the growth and decline of a variant or clade. Different genotypes are phylogenetically related and the molecular clock constrains how the τ_i change along the tree. For simplicity, we will models this as a Gaussian as well. Referring to the parent of i as p_i, we have for the full log-LH

$$\mathcal{L} = \sum_i \left(\frac{(\mu(\tau_i - \tau_{p_i}) - d_i)^2}{2(d_i+1)} + \sum_{\alpha \in s_i} \frac{(t_\alpha - \tau_i)^2}{2\sigma^2} \right) \tag{2}$$

where d_i is the number of mutations between i and p_i. We want to optimize this with respect to the genotype timings τ_i. Diffentiating with respect to τ_k, we have

$$\partial_{\tau_k} \mathcal{L} = \mu \frac{(\mu(\tau_k - \tau_{p_k}) - d_k)}{d_k + 1} + \sum_{\alpha \in s_k} \frac{(\tau_k - t_\alpha)}{\sigma^2} - \sum_{c \in k} \mu \frac{(\mu(\tau_c - \tau_k) - d_c)}{d_c + 1} = 0 \tag{3}$$

This is a sparse linear system that can be readily solved for τ_k. This could also be solved analytically in a forward backward fashion. It will be useful to define the average time $\bar{t}_i$ and the number of observations of genotype i as n_i to simplify the above to

$$\partial_{\tau_k} \mathcal{L} = \mu \frac{(\mu(\tau_k - \tau_{p_k}) - d_k)}{d_k + 1} + n_k \frac{(\tau_k - \bar{t}_k)}{\sigma^2} - \mu \sum_{c \in k} \frac{(\mu(\tau_c - \tau_k) - d_c)}{d_c + 1} = 0 \tag{4}$$

The resulting times can then be plugged into $\mathcal{L}$, and we can optimize σ and maybe μ. Once those are optimized, we can compare each node sampling time to its distribution. If we did the forward-backward distributions, we could also look at the leave-on-out signal.

Forward-backward solution

For a terminal node k, the optimal position given the position of the parent τ_{p_k} is

$$\tau_k = \left(\frac{n_k}{\sigma^2} + \frac{\mu^2}{d_k + 1} \right)^{-1} \left(\frac{n_k \bar{t}_k}{\sigma^2} + \mu \frac{\mu \tau_{p_k} + d_k}{d_k + 1} \right) = a + b\tau_{p_k}. \tag{5}$$

This expression weighs the evidence of the node being placed close to the average samples against the position of and the mutations relative to the parent. A similar calculation can be done for internal nodes where we have additional contributions from the children where we plug in the expression for their optimal position given the parent.

$$\mu \frac{(\mu(\tau_k - \tau_{p_k}) - d_k)}{d_k + 1} + n_k \frac{(\tau_k - \bar{t}_k)}{\sigma^2} - \mu \sum_{c \in k} \frac{(\mu(a_c + b_c \tau_k - \tau_k) - d_c)}{d_c + 1} = 0 \tag{6}$$

Collecting all terms proportional to τ_k, we find

$$\tau_k\left(\frac{n_k}{\sigma^2}+\frac{\mu^2}{d_k+1}+\mu^2\sum_{c\in k}\frac{1-b_c}{d_c+1}\right)=\frac{n_k\bar{t}_k}{\sigma^2}+\mu\frac{\mu\tau_{p_k}+d_k}{d_k+1}+\mu\sum_{c\in k}\frac{\mu a_c-d_c}{d_c+1} \tag{7}$$

Which can again be solved for τ_k and is a linear function of τ_{p_k}.

$$\tau_k=\left(\frac{n_k}{\sigma^2}+\frac{\mu^2}{d_k+1}+\mu^2\sum_{c\in k}\frac{1-b_c}{d_c+1}\right)^{-1}\left(\frac{n_k\bar{t}_k}{\sigma^2}+\mu\frac{\mu\tau_{p_k}+d_k}{d_k+1}+\mu\sum_{c\in k}\frac{\mu a_c-d_c}{d_c+1}\right)=a+b\tau_{p_k} \tag{8}$$

At the root, the term from the parent is absent and their is no conditioning on τ_{p_k} anymore

$$\tau_k=\left(\frac{n_k}{\sigma^2}+\mu^2\sum_{c\in k}\frac{1-b_c}{d_c+1}\right)^{-1}\left(\frac{n_k\bar{t}_k}{\sigma^2}+\mu\sum_{c\in k}\frac{\mu a_c-d_c}{d_c+1}\right)=a+b\tau_{p_k} \tag{9}$$

Once all the root τ is known, all other τ_k can be calculated in one backward pass.

The optimal timings can be calculated in linear time, along with the cost-function. This cost-function can then be optimized for σ and μ.

Optimizing σ and μ

Differentiating the log-LH with respect to μ, we have

$$\frac{\partial\mathcal{L}}{\partial\mu}=\sum_i\frac{\mu(\tau_i-\tau_{p_i})^2-(\tau_i-\tau_{p_i})d_i}{2(d_i+1)}=\sum_i\frac{\mu(\tau_i-\tau_{p_i})^2-(\tau_i-\tau_{p_i})d_i}{2(d_i+1)}=\mu\sum_i\frac{(\tau_i-\tau_{p_i})^2}{2(d_i+1)}-\sum_i\frac{(\tau_i-\tau_{p_i})d_i}{2(d_i+1)}=0 \tag{10}$$

and therefore

$$\mu=\frac{\sum_i\frac{(\tau_i-\tau_{p_i})d_i}{(d_i+1)}}{\sum_i\frac{(\tau_i-\tau_{p_i})^2}{(d_i+1)}} \tag{11}$$

The values of τ of course depend on the choice of μ and this problem therefore has to be solved iteratively.

The optimization with respect to σ does not work as naively expected. Iteration drives σ to small values and tends to result in underestimation of the evolutionary rate. This phenomenon is similar to the issues observed in the Gaussian tree-regression that requires large tip variances. My current thinking is that small σ or small tip variance results in "local fits" where similar sequences that were sampled several months apart drive down the rate estimate.

The second derivative of the rate μ, which is the inverse confidence of the rate estimate, is simply

$$\sum_i\frac{(\tau_i-\tau_{p_i})^2}{2(d_i+1)} \tag{12}$$

and the error in μ

$$\Delta\mu=\left(\sum_i\frac{(\tau_i-\tau_{p_i})^2}{2(d_i+1)}\right)^{-1/2} \tag{13}$$

Coalesence prior

One reason for the underestimation of the rate could be that the coalescence prior is not properly accounted for. Coalescence will shorten branches and therefore increase the rate estimate. The coalescence prior is proportional to the contemporary lineages. But in principle a linear cost on branch length should be possible to implement. Augmenting the negative log-LH with a term that accounts for coalescence, yields

$$\mathcal{L}=\sum_i\left(\frac{(\mu(\tau_i-\tau_{p_i})-d_i)^2}{2(d_i+1)}+\int_{\tau_{p_i}}^{\tau_i}\frac{k(t)-1}{2T_c}dt+\sum_{\alpha\in s_i}\frac{(t_\alpha-\tau_i)^2}{2\sigma^2}\right) \tag{14}$$

Here we ignored the non-exponential pre-factor which should introduce a term $n \log T_c$ similar to the normalization of the Gaussians. If we differentiate $\mathcal{L}$ with respect to τ_k, we pick up terms that correspond to the differential of the upper and lower limit of the integral.

$$\partial_{\tau_k} \mathcal{L} = \mu \frac{(\mu(\tau_k - \tau_{p_k}) - d_k)}{d_k + 1} + \gamma_k(|k| - 1) + n_k \frac{(\tau_k - \bar{t}_k)}{\sigma^2} - \mu \sum_{c \in k} \frac{(\mu(\tau_c - \tau_k) - d_c)}{d_c + 1} \tag{15}$$

where $\gamma_k = (k_k - 1)/T_c$ is the coalescence rate at node k and $|k|$ is the number of children of k. The recursion relations above can be readily adapted to include the coalescence term: Wherever we have a τ_k in the numerator, we have to subtract $\gamma_k(|k| - 1)$.

Again, this is a procedure that needs to be done iteratively since the coalescence rate depends on the values of τ across the tree.

strictly speaking, the coalescence terms from the parent branch and the child branches have slightly different weights:

$$\gamma_k(|k| - 1) = \frac{|k| m_k - (m_k - 1)}{T_c} = \frac{(|k| - 1) m_k + 1}{T_c} = \frac{(|k| - 1) m_k}{T_c} + \frac{1}{T_c} \tag{16}$$

where the last term $\frac{1}{T_c}$ is accounts for the difference in rate before and after the merger.