9.3* ML estimation of the species tree for three species given the gene tree at one locus, with one sequence from each species. Use the gene tree $G = ((A, B), C)$, with node ages t_0 and t_1, of Figure 9.22a as given data to evaluate the likelihood for the species

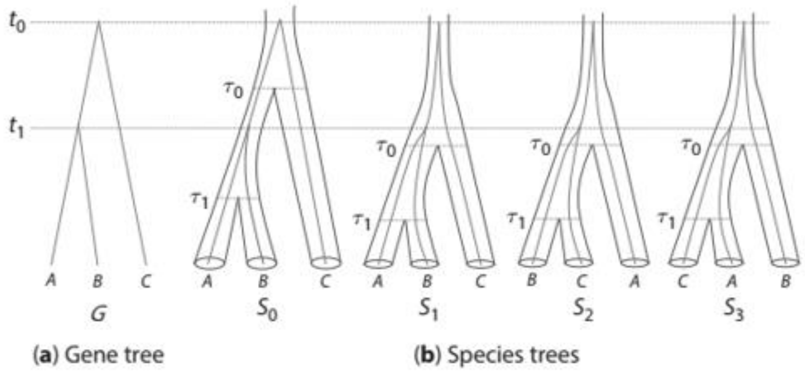

Fig. 9.22 Estimation of the species tree for three species using a gene tree for one locus, with one sequence from each species. (**a**) The gene tree, with topology $G = ((A, B), C)$ and node ages t_0 and t_1, is the given data. (**b**) The species trees with their parameters. Species tree S_0 involves parameters τ_0 and τ_1, under the constraints $\tau_1 \le t_1 \le \tau_0 \le t_0$, while each of species trees S_1, S_2, and S_3 involves parameters τ_0 and τ_1, with $\tau_1 \le \tau_0 \le t_1 \le t_0$. Each of the species tree also involve two population size parameters θ_0 and θ_1, which are not shown. The Maximum Tree algorithm assumes $\theta_0 = \theta_1 = \theta$. Note that species trees S_0 and S_1 have the same topology.

trees of Figure 9.22b, under the assumption that all populations have the same θ. Treat species trees S_0 and S_1 separately even though they have the same tree topology. Show that the ML estimate of the species tree is S_0, with $\hat{\tau}_0 = t_0$ and $\hat{\tau}_1 = t_1$. [Hint: Write down the likelihood function for species tree S_0, which is the multispecies coalescent density for the gene tree, $f(G, t_0, t_1 | S_0, \tau_0, \tau_1, \theta)$, and maximize it by adjusting τ_0, τ_1, and θ under the constraints $\tau_1 \le t_1 \le \tau_0 \le t_0$. Then repeat the analysis for species trees S_1, S_2, and S_3.]

Solution.

According to Eq. (9.45) in (Yang 2014a), and because $\theta_0 = \theta_1 = \theta$, we have

$$f(G, t | S, \Theta) = \left(\frac{2}{\theta}\right)^C e^{-\frac{2}{\theta}T},$$

where C is the number of coalescent events on the gene tree, T is the so-called "total per-lineage-pair coalescent time" summed over all populations and all gene trees, and $\Theta = (S, \tau_0, \tau_1, \theta)$. Denote the "total per-lineage-pair coalescent time" for species tree S_k as T_k. According to the statement of the problem, we have $C = 2$ for all species trees, and

$$T_0 = (t_1 - \tau_1) + (t_0 - \tau_0),$$

$$T_1 = T_2 = T_3 = (\tau_0 - \tau_1) + 3(t_1 - \tau_0) + (t_0 - t_1).$$

The logic is to calculate the maximum likelihood under four species trees S_0, S_1, S_2, S_3 one by one and compare their values.

a)

As to the species trees S_1, S_2, S_3, we have

$$f(G, t|S_k, \tau_0, \tau_1, \theta) = \left(\frac{2}{\theta}\right)^2 e^{-\frac{2}{\theta}T_1} = \left(\frac{2}{\theta}\right)^2 e^{-\frac{2}{\theta}((\tau_0-\tau_1)+3(t_1-\tau_0)+(t_0-t_1))}$$

where $k = 1,2,3$. Thus, we are looking for

$$(\hat{\tau}_0, \hat{\tau}_1, \hat{\theta}) = \underset{\tau_0,\tau_1,\theta}{\operatorname{argmax}} \left\{\left(\frac{2}{\theta}\right)^2 e^{-\frac{2}{\theta}((\tau_0-\tau_1)+3(t_1-\tau_0)+(t_0-t_1))} | \tau_1 \le \tau_0 \le t_1 \le t_0, \theta > 0\right\}.$$

Rewrite the likelihood function as

$$\begin{aligned} f(G, t|S_k, \tau_0, \tau_1, \theta) &= \left(\frac{2}{\theta}\right)^2 e^{-\frac{2}{\theta}((\tau_0-\tau_1)+3(t_1-\tau_0)+(t_0-t_1))} \\ &= \left(\frac{2}{\theta}\right)^2 e^{-\frac{2}{\theta}(2t_1+t_0-2\tau_0-\tau_1)}. \end{aligned}$$

Define $T^* = 2t_1 + t_0 - 2\tau_0 - \tau_1$.

Set

$$\frac{\partial\left(\left(\frac{2}{\theta}\right)^2 e^{-\frac{2}{\theta}T^*}\right)}{\partial\theta} = 0,$$

and by solving the above, we obtain $\hat{\theta} = T^*$.

Thus, the maximum of $f(G, t|S_k, \tau_0, \tau_1, \theta) = \left(\frac{2}{\theta}\right)^2 e^{-\frac{2}{\theta}(2t_1+t_0-2\tau_0-\tau_1)}$ can be achieved when $\theta = T^* = 2t_1 + t_0 - 2\tau_0 - \tau_1$. Accordingly, we have

$$f(G, t|S_k, \tau_0, \tau_1, \theta = T^*) = \left(\frac{2}{2t_1 + t_0 - 2\tau_0 - \tau_1}\right)^2 e^{-2}.$$

Because of the constraint $\tau_1 \le \tau_0 \le t_1 \le t_0$, it can be seen that the maximum is achieved when $\tau_0 = \tau_1 = t_1$. Hence, we have

$$\begin{aligned} &\max\left\{\left(\frac{2}{\theta}\right)^2 e^{-\frac{2}{\theta}((\tau_0-\tau_1)+3(t_1-\tau_0)+(t_0-t_1))} \middle| \tau_1 \le \tau_0 \le t_1 \le t_0, \theta > 0\right\} \\ &= f(G, t|S_k, \tau_0 = t_1, \tau_1 = t_1, \theta = T^*) \\ &= \left(\frac{2}{t_0 - t_1}\right)^2 e^{-2}, \end{aligned}$$

where $k = 1,2,3$.

b)

As to species tree S_0,

$$\left(\frac{2}{\theta}\right)^2 e^{-\frac{2}{\theta}T_0} = \left(\frac{2}{\theta}\right)^2 e^{-\frac{2}{\theta}((t_1-\tau_1)+(t_0-\tau_0))},$$

s.t. $\tau_1 \le t_1 \le \tau_0 \le t_0, \theta > 0$.

In other words, we are looking for

$$\left(\hat{\tau}_0, \hat{\tau}_1, \hat{\theta}\right) = \underset{\tau_0,\tau_1,\theta}{\operatorname{argmax}} \left\{ \left(\frac{2}{\theta}\right)^2 e^{-\frac{2}{\theta}\left((t_1-\tau_1)+(t_0-\tau_0)\right)} | \tau_1 \leq t_1 \leq \tau_0 \leq t_0, \theta > 0 \right\}.$$

According to a), it is already known that the maximum of $f(x)$ is achieved when $\hat{\theta} = T^* = (t_1 - \tau_1) + (t_0 - \tau_0)$. Considering the constraint $\tau_1 \leq t_1 \leq \tau_0 \leq t_0$, it is obvious that setting $\tau_0 = t_0, \tau_1 = t_1, \theta = (t_1 - \tau_1) + (t_0 - \tau_0) = 0$, the maximum of the likelihood function is achieved at $\left(\frac{2}{\theta}\right)^2 e^{-\frac{2}{\theta}\left((t_1-t_1)+(t_0-t_0)\right)} = \left(\frac{2}{\theta}\right)^2 \to \infty$.

Based on a) and b), when $S = S_0, \tau_0 = t_0, \tau_1 = t_1, \theta = (t_1 - \tau_1) + (t_0 - \tau_0) = 0$, the maximum of the likelihood is achieved.