Notation: Given a segment length ℓ and a number m of heterozygous sites, we want to find t, the age of the segment in generations. We say ℓ is measured in Morgans and ℓ' in base-pairs. We assume a fixed demography of size N_e and a pairwise coalscence rate $\gamma = \frac{1}{N_e}$.

We write $\rho = 2\ell$ and $\mu = 2c * \ell'$, where c is the fixed mutation rate.

Maximum likelihood

Here, we maximise the likelihood function $l(\ell, m|t)$.

Using only ℓ

If we ignore m we can write $l(\ell|t) = te^{-t\rho}$, which has derivative $e^{-t\rho}(1 - t\rho)$ with respect to t, and a maximum at $t = \frac{1}{\rho}$.

Using ℓ and m

We assume mutation arise following a Poisson($t\mu$) process. To simplify notation, μ, like ρ, is a function of segment length. This leads to

$$l(\ell, m|t) = l(\ell|t)l(m|\ell, t) = te^{-t\rho}\frac{(t\mu)^m}{m!}e^{-t\mu}.$$

Taking the derivative w.r.t t, we get

$$e^{-t(\mu+\rho)}t^m((m+1) - t(\rho+\mu))$$

times a constant, which gives a maximum at $t = \frac{m+1}{(\rho+\mu)}$.

Bayes

Here we also introduce a prior on t, namely $\pi(t) \sim \text{Exp}(\gamma)$ and try to compute $\mathbb{E}[t|\ell, \mu]$.

Using only ℓ

As before, we can start by ignoring mutations to get

$$\mathbb{E}[t|\ell] = \int_0^\infty tp(t|\ell)\mathrm{d}t = \int_0^t t\frac{\pi(t)p(\ell|t)}{p(l)}\mathrm{d}t.$$

We already know $\pi(t) = \gamma e^{-\gamma t}$ and $p(\ell|t) = te^{-t\rho}$, so we can compute the integral

$$\int_0^\infty tp(t)p(\ell|t)\mathrm{d}t = \int_0^\infty \gamma t^2 e^{-t(\rho+\gamma)} = \frac{2\gamma}{(\rho+\gamma)^3}.$$

Moreover,

$$p(l) = \int_0^\infty p(t)p(\ell|t)\mathrm{d}t = \int_0^\infty \gamma t e^{-t(\rho+\gamma)} = \frac{\gamma}{(\rho+\gamma)^2}.$$

This gives

$$\mathbb{E}[t|\ell] = \frac{2\gamma/(\rho+\gamma)^3}{\gamma/(\rho+\gamma)^2} = \frac{2}{\rho+\gamma}.$$

Using ℓ and m

Following a similar strategy, we find

$$\mathbb{E}[t|\ell, m] = \frac{\int_0^\infty tp(t)p(m|\ell, t)p(\ell|t)\mathrm{d}t}{p(l)p(m|l)}.$$

For the integral, we get

$$\int_0^\infty t\gamma e^{-t\gamma} te^{-t\rho}\frac{(t\mu)^m}{m!}e^{-t\mu}\mathrm{d}t = \frac{\gamma\mu^m}{m!}\int_0^\infty t^{m+2}e^{-t(\gamma+\rho+\mu)}\mathrm{d}t = \frac{\gamma\mu^m}{m!}\frac{(m+2)!}{(\gamma+\rho+\mu)^{m+3}},$$

and for the denominator, we only need

$$p(m|l) = \int_0^\infty p(m|l, t)p(t) = \frac{\gamma\mu^m}{(\gamma+\mu)^{m+1}}.$$

This gives

$$\frac{\gamma\mu^m(m+1)}{(\gamma+\rho+\mu)^{m+2}}$$

which gives final result

$$\frac{(m+2)}{\gamma+\rho+\mu}.$$

The Erlang

Romain made a good observation: The segment might be longer than what we observe. So it might be better to use a "segment of length at least u" estimator.

For constant demography, this is pretty easy, since $p(t|l \geq u)$ is Erlang-2 with rate $(2l+\gamma)$. But can we mix mutations in there? Yes! Because we in our case $p(m|t, l > u) = p(m|t, l)$, because we only care about observable mutations. This gives

$$\mathbb{E}[t|m, l \geq u] = \frac{\int_0^\infty tp(t)p(m|l > u, t)p(l > u|t)\mathrm{d}t}{p(l > u)p(m|l > u)}.$$

We tease this apart to get

$$\int_0^\infty t\gamma e^{-t\gamma}\frac{(t\mu)^m}{m!}e^{-t\mu}2te^{-2t\rho}\mathrm{d}t = 2\frac{\gamma\mu^m}{m!}\int_0^\infty t^{m+2}e^{-t(\gamma+2\rho+\mu)}\mathrm{d}t = 2\frac{\gamma\mu^m}{m!}\frac{(m+2)!}{(\gamma+2\rho+\mu)^{m+3}},$$

using $p(l > u|t) = 2te^{-2t\rho}$, where ρ is the generation-wise recombination rate on u. And I think that's it, really. We get

$$\mathbb{E}[t|m, l \geq u] = \frac{2(m+2)}{\gamma+2\rho+\mu}.$$

Variable-size demography

Notation here: $0 = T_0, T_1, \ldots, T_K = \infty$ define the time bins and we write $\Delta_j = T_j - T_{j-1}$, such that $p(t) = \gamma_k e^{-\sum_{j=1}^{k-1}\Delta_j\gamma_j - (t-T_k)\gamma_k}$.

The numerator in $\mathbb{E}[t|m, l \geq u]$ then becomes

$$\begin{aligned}
&\sum_{k=1}^{K} \int_{T_{k-1}}^{T_k} tp(t)p(m|u,t)p(l > u|t)\mathrm{d}t \\
&= \sum_{k=1}^{K} \gamma_k e^{-\sum_{j=1}^{k-1} \Delta_j \gamma_j + T_k \gamma_k} 2\gamma_k \frac{\mu^m}{m!} \int_{T_{k-1}}^{T_k} t^{m+2} e^{-t(\gamma_k + 2\rho + \mu)} \mathrm{d}t \\
&= \sum_{k=1}^{K} \gamma_k e^{-\sum_{j=1}^{k-1} \Delta_j \gamma_j + T_k \gamma_k} 2\gamma_k \frac{\mu^m}{m!} \frac{(m+3)!}{(\gamma_k + 2\rho + \mu)^{m+3}} \int_{T_{k-1}}^{T_k} \frac{(\gamma_k + 2\rho + \mu)^{m+3}}{(m+3)!} t^{m+2} e^{-t(\gamma_k + 2\rho + \mu)} \mathrm{d}t \\
&= \sum_{k=1}^{K} \gamma_k e^{-\sum_{j=1}^{k-1} \Delta_j \gamma_j + T_k \gamma_k} 2\gamma_k \frac{\mu^m}{m!} \frac{1}{(\gamma_k + 2\rho + \mu)^{m+3}} \times \\
&\qquad \Big[\gamma\big(m+3, (\gamma_k + 2\rho + \mu) T_{k+1}\big) - \gamma\big(m+3, (\gamma_k + 2\rho + \mu) T_k\big) \Big],
\end{aligned}$$

where $\gamma(\cdot)$ is one of the incomplete gamma function available in e.g. boost. We can evaluate this in $O(K)$ time.

Threading to multiple sequences

Suppose we want to thread to a genealogy on N leaves. Close to 0, the new sequence coalesces with rate $N\gamma$, decreasing to γ as $t \to \inf$. This decrease depends on a complicated convolution of $\binom{k}{2}\gamma$-rate variables, for $k = N, \ldots, 1$. That is, the new sequence coalesces with rate $k(t)\gamma$, where $k(t)$ is a random variable expressing the number of open lineages at time t.

We're going to assume a piecewise-constant demography and we're going to derive an approximation for $V_j := \int_{T_{j-1}}^{T_j} k(t)\mathrm{d}t$. Denote the coalescence time of the new sequence by X_{N+1}. We can recursively write

$$\mathbb{E}[V_{N+1,j}] = V_{N,j} + \mathbb{P}(X_{N+1} \geq T_j) + \Big(\mathbb{E}[X_{N+1} \,|\, T_{j-1} \leq X_{N+1} < T_j] - T_{j-1} \Big),$$

where

$$\mathbb{P}(X_{N+1} \geq T_j) = e^{-\sum_{i=1}^{j} V_j \gamma_j},$$

and we do an approximation where we say X_{N+1} converges with a fixed rate $\omega_{N,j} := \gamma_j \frac{V_{N,j}}{T_j - T_{j-1}}$ in the j-th interval. This gives a closed-form expression

$$\mathbb{E}[X_{N+1} \,|\, T_{j-1} \leq X_{N+1} < T_j] \approx \frac{1}{\int_{T_{j-1}}^{T_j} \omega_{N,j} e^{-t\omega_{N,j}} \mathrm{d}t} \int_{T_{j-1}}^{T_j} t\omega_{N,j} e^{-t\omega_{N,j}} \mathrm{d}t,$$

which is easy and we'll write down when we get the time. We say $V_{1,j} = 1$ for all j.

This approximation will slightly overestimate the tree volume due to the assumption that coalescence occurs exponentially within time intervals, while in reality it's more biased toward the bottom. This is hopefully not too bad, and might go away with really small intervals or some hacks.

Applying this approximation to the threading routine, we can say that between T_{j-1} and T_j, the new sequence coalesces with rate $\gamma_j \omega_{N,j}$. This means we can replace all γ_j in the big equation above with $\gamma_j \omega_{N,j}$.

Why is this good? It means we can use any piecewise constant demography and condition on the size of the ARG we're threading to in $O(K)$ time by iteratively updating the demography.