Probabilities for Y-Chromosome Markers Matches

L. David Roper, roperld@vt.edu

The paper "Estimating the Time to the MRCA for the Y chromosome or mtDNA for a Pair of Individuals" by Bruce Walsh, Univ. of Arizona (http://nitro.biosci,arizona.edu/zdownload/current_ms/MCRA.pdf) contains equations for probabilities versus time in generations to the Most Recent Common Ancestor for a pair of individuals.

The probability versus time in generations (t) when measuring n markers with no mismatches:

$p_n(t) = 2n\mu \exp(-2n\mu t)$.

This function is $2n\mu$ at 0 and drops to 0 at ∞.

The probability versus time in generations (t) when measuring n markers with k matches:

$$p_{nk}(t) = \frac{\prod_{i=0}^{n-k}[2\mu(n-i)]}{2^{n-k}(n-k)!\mu^{n-k}} \frac{(1-\exp[-2\mu t])^{n-k}}{\exp[2\mu k t]}.$$

These functions start at 0 for t=0 and peak and then fall to 0 at ∞. To find the value of p at the peak, set the first derivative to 0:

$$\frac{dp}{dt} = 4(-1)^{n-k}\mu^2(1-e^{-2\mu t})^{n-k} \frac{e^{-2\mu t-2t\mu k}n - e^{-2t\mu k}k}{-1+e^{-2\mu t}} \frac{\Gamma(-k+1)}{\Gamma(-n)\Gamma(n-k+1)} = 0.$$

The solution is: $t_p = \frac{1}{2}\frac{\ln\frac{n}{k}}{\mu}$ or $2\mu t_p = \ln\frac{n}{k}$.

Examples for $\mu = \frac{1}{500}$:

n=12:

- k=11: $t_p = 250\ln\frac{12}{11} = 21.8$
- k=10: $t_p = 250\ln\frac{12}{10} = 45.6$
- k=9: $t_p = 250\ln\frac{12}{9} = 71.9$

n=25:

- k=24: $t_p = 250\ln\frac{25}{24} = 10.21$
- k=23: $t_p = 250\ln\frac{25}{23} = 20.8$
- k=22: $t_p = 250\ln\frac{25}{22} = 32.0$

n=50:

- k=49: $t_p = 250\ln\frac{50}{49} = 5.1$
- k=48: $t_p = 250\ln\frac{50}{48} = 10.2$
- k=47: $t_p = 250\ln\frac{50}{47} = 15.5$

Then the value at the peak is:

$$p(t_p) = \frac{\prod_{i=0}^{n-k}[2\mu(n-i)]}{2^{n-k}(n-k)!\mu^{n-k}} \frac{\left(1-\exp\left[-\ln\frac{n}{k}\right]\right)^{n-k}}{\exp\left[k\ln\frac{n}{k}\right]} = \frac{\prod_{i=0}^{n-k}[2\mu(n-i)]}{2^{n-k}(n-k)!\mu^{n-k}} \frac{\left(1-\frac{k}{n}\right)^{n-k}}{\exp\left[k\ln\frac{n}{k}\right]}.$$

The two MRCA values at one-half the peak value are given by:

$$p\left(t_{\frac{1}{2}}\right) = \frac{p(t_p)}{2}.$$

This is difficult to solve analytical, but we can solve it for specific examples of interest for $\mu = \frac{1}{500}$:

n=12:

k=11: $t_{\frac{1}{2}} = 5.1$ and 58.4

n=25:

k=24: $t_{\frac{1}{2}} = 2.4$ and 27.3

n=50:

k=49: $t_{\frac{1}{2}} = 1.2$ and 13.5

Now plot the curves for specific examples of interest for $\mu = \frac{1}{500}$:

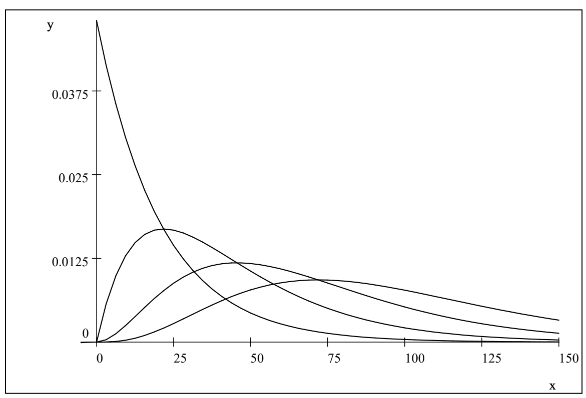
n=12; k=11, 10 & 9:

$$p_{12}(t) = 2 * 12 * \frac{1}{500}\exp(-2 * 12 * \frac{1}{500}t) = \frac{6}{125}e^{-\frac{6}{125}t}$$

$$p_{12,11}(t) = \frac{\prod_{i=0}^{12-11}\left[2\left(\frac{1}{500}\right)(12-i)\right]}{2^{12-11}(12-11)!\left(\frac{1}{500}\right)^{12-11}}\frac{\left(1-\exp\left[-2\left(\frac{1}{500}\right)t\right]\right)^{12-11}}{\exp\left[2\left(\frac{1}{500}\right)11t\right]} = 0.528\,\frac{1.0-1.0\exp(-0.004\,t)}{\exp(0.044\,t)}$$

$$p_{12,10}(t) = \frac{\prod_{i=0}^{12-10}\left[2\left(\frac{1}{500}\right)(12-i)\right]}{2^{12-10}(12-10)!\left(\frac{1}{500}\right)^{12-10}}\frac{\left(1-\exp\left[-2\left(\frac{1}{500}\right)t\right]\right)^{12-10}}{\exp\left[2\left(\frac{1}{500}\right)10t\right]} = 2.64\,\frac{(1.0-1.0\exp(-0.004\,t))^2}{\exp(0.04t)}$$

$$p_{12,9}(t) = \frac{\prod_{i=0}^{12-9}\left[2\left(\frac{1}{500}\right)(12-i)\right]}{2^{12-9}(12-9)!\left(\frac{1}{500}\right)^{12-9}}\frac{\left(1-\exp\left[-2\left(\frac{1}{500}\right)t\right]\right)^{12-9}}{\exp\left[2\left(\frac{1}{500}\right)9t\right]} = 7.92\,\frac{(1.0-1.0\exp(-0.004\,t))^3}{\exp(0.036\,t)}$$

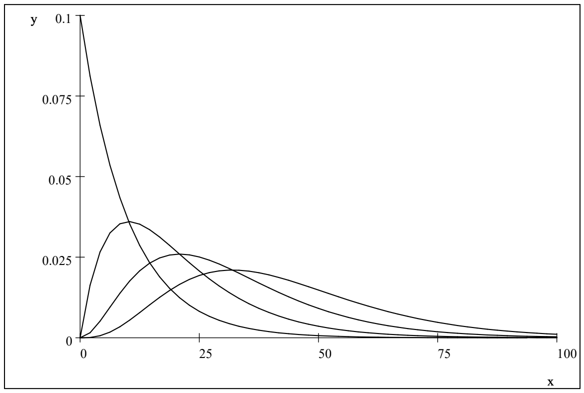


$$p_{25}(t) = 2 * 25 * \frac{1}{500}\exp(-2 * 25 * \frac{1}{500}t) = \frac{1}{10}e^{-\frac{1}{10}t}$$

$$p_{25,24}(t) = \frac{\prod_{i=0}^{25-24}\left[2\left(\frac{1}{500}\right)(25-i)\right]}{2^{25-24}(25-24)!\left(\frac{1}{500}\right)^{25-24}}\frac{\left(1-\exp\left[-2\left(\frac{1}{500}\right)t\right]\right)^{25-24}}{\exp\left[2\left(\frac{1}{500}\right)24t\right]} : \frac{2.4}{e^{0.096\,t}}(1.0 - 1.0e^{-0.004\,t})$$

$$p_{25,23}(t) = \frac{\prod_{i=0}^{25-23}\left[2\left(\frac{1}{500}\right)(25-i)\right]}{2^{25-23}(25-23)!\left(\frac{1}{500}\right)^{25-23}}\frac{\left(1-\exp\left[-2\left(\frac{1}{500}\right)t\right]\right)^{25-23}}{\exp\left[2\left(\frac{1}{500}\right)23t\right]} : \frac{27.6}{e^{0.092\,t}}(1.0 - 1.0e^{-0.004\,t})^2$$

$$p_{25,22}(t) = \frac{\prod_{i=0}^{25-22}\left[2\left(\frac{1}{500}\right)(25-i)\right]}{2^{25-22}(25-22)!\left(\frac{1}{500}\right)^{25-22}} \frac{\left(1-\exp\left[-2\left(\frac{1}{500}\right)t\right]\right)^{25-22}}{\exp\left[2\left(\frac{1}{500}\right)22t\right]} : \frac{202.4}{e^{0.088t}}\left(1.0 - 1.0e^{-0.004t}\right)^3$$

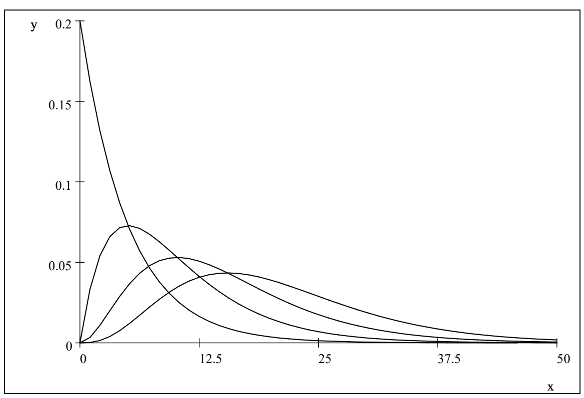


$$p_{50}(t) = 2 * 50 * \frac{1}{500}\exp(-2 * 50 * \frac{1}{500}t) = \frac{1}{5}e^{-\frac{1}{5}t}$$

$$p_{50,49}(t) = \frac{\prod_{i=0}^{50-49}\left[2\left(\frac{1}{500}\right)(50-i)\right]}{2^{50-49}(50-49)!\left(\frac{1}{500}\right)^{50-49}} \frac{\left(1-\exp\left[-2\left(\frac{1}{500}\right)t\right]\right)^{50-49}}{\exp\left[2\left(\frac{1}{500}\right)49t\right]} = 9.8\frac{1.0-1.0\exp(-0.004t)}{\exp(0.196t)}$$

$$p_{50,48}(t) = \frac{\prod_{i=0}^{50-48}\left[2\left(\frac{1}{500}\right)(50-i)\right]}{2^{50-48}(50-48)!\left(\frac{1}{500}\right)^{50-48}} \frac{\left(1-\exp\left[-2\left(\frac{1}{500}\right)t\right]\right)^{50-48}}{\exp\left[2\left(\frac{1}{500}\right)48t\right]} = 235.2\frac{(1.0-1.0\exp(-0.004t))^2}{\exp(0.192t)}$$

$$p_{50,47}(t) = \frac{\prod_{i=0}^{50-47}\left[2\left(\frac{1}{500}\right)(50-i)\right]}{2^{50-47}(50-47)!\left(\frac{1}{500}\right)^{50-47}} \frac{\left(1-\exp\left[-2\left(\frac{1}{500}\right)t\right]\right)^{50-47}}{\exp\left[2\left(\frac{1}{500}\right)47t\right]} = 3684.8\frac{(1.0-1.0\exp(-0.004t))^3}{\exp(0.188t)}$$



Note in the three plots that each curve crosses the peak of the succeeding curve. We now prove this analytically:

Proof that k curve crosses peak of k-1 curve:

$$p(t) = \frac{\prod_{i=0}^{n-k}[2\mu(n-i)]}{2^{n-k}(n-k)!\mu^{n-k}} \frac{(1-\exp[-2\mu t])^{n-k}}{\exp[2\mu k t]}$$

$t(\text{k peak}) = \frac{1}{2}\frac{\ln\frac{n}{k}}{\mu}$ or $2\mu t = \ln\frac{n}{k}$

t(k-1 peak) $= \frac{1}{2}\frac{\ln\frac{n}{k-1}}{\mu}$ or $2\mu t = \ln\frac{n}{k-1}$

Peak of k curve:

$$p(\text{k peak}) = \frac{\prod_{i=0}^{n-k}[2\mu(n-i)]}{2^{n-k}(n-k)!\mu^{n-k}}\frac{\left(1-\exp\left[-\ln\frac{n}{k}\right]\right)^{n-k}}{\exp\left[k\ln\frac{n}{k}\right]} = \frac{\prod_{i=0}^{n-k}[2\mu(n-i)]}{2^{n-k}(n-k)!\mu^{n-k}}\frac{\left(1-\frac{k}{n}\right)^{n-k}}{\exp\left[k\ln\frac{n}{k}\right]}$$

Peak of k-1 curve:

$$p(\text{k-1 peak}) = \frac{\prod_{i=0}^{n-k+1}[2\mu(n-i)]}{2^{n-k+1}(n-k+1)!\mu^{n-k+1}}\frac{\left(1-\frac{k-1}{n}\right)^{n-k+1}}{\exp\left[(k-1)\ln\frac{n}{k-1}\right]}$$

$$= \frac{(k-1)\prod_{i=0}^{n-k}[2\mu(n-i)]}{(n-k+1)2^{n-k}(n-k)!\mu^{n-k}}\frac{\left(1-\exp\left[-\ln\frac{n}{k-1}\right]\right)^{n-k}\left(1-\frac{k-1}{n}\right)}{\exp\left[k\ln\frac{n}{k-1}\right]\exp\left[-\ln\frac{n}{k-1}\right]} = \frac{(k-1)\left(1-\frac{k-1}{n}\right)}{(n-k+1)\frac{k-1}{n}}\frac{\prod_{i=0}^{n-k}[2\mu(n-i)]}{2^{n-k}(n-k)!\mu^{n-k}}\frac{\left(1-\frac{k-1}{n}\right)^{n-k}}{\exp\left[k\ln\frac{n}{k-1}\right]}$$

k curve at peak of k-1 curve:

$$p(\text{k curve at k-1 peak}) = \frac{\prod_{i=0}^{n-k}[2\mu(n-i)]}{2^{n-k}(n-k)!\mu^{n-k}}\frac{\left(1-\frac{k-1}{n}\right)^{n-k}}{\exp\left[k\ln\frac{n}{k-1}\right]}$$

But:

$$\frac{(k-1)\left(1-\frac{k-1}{n}\right)}{(n-k+1)\frac{k-1}{n}} = 1$$

Therefore, p(k-1 peak) $= p$(k curve at k-1 peak) Q.E.D

Proof that n curve with no mismatches crosses peak of k=n-1 curve:

Peak of k curve:

$$p(\text{k peak}) = \frac{\prod_{i=0}^{n-k}[2\mu(n-i)]}{2^{n-k}(n-k)!\mu^{n-k}}\frac{\left(1-\frac{k}{n}\right)^{n-k}}{\exp\left[k\ln\frac{n}{k}\right]}$$

Peak of k=n-1 curve:

$$p(\text{k=n-1 peak}) = \frac{\prod_{i=0}^{1}[2\mu(n-i)]}{2^{1}(1)!\mu^{1}}\frac{\left(1-\frac{n-1}{n}\right)^{1}}{\exp\left[(n-1)\ln\frac{n}{n-1}\right]} = \frac{2\mu n 2\mu(n-1)}{2\mu}\frac{\left(1-\frac{n-1}{n}\right)^{1}}{\exp\left[(n-1)\ln\frac{n}{n-1}\right]} =$$

$$2\mu n(n-1)\frac{1-\frac{n-1}{n}}{\exp\left((n-1)\ln\frac{n}{n-1}\right)} = \frac{2\mu(n-1)}{\exp\left((n-1)\ln\frac{n}{n-1}\right)} = 2\mu(n-1)\exp\left((1-n)\ln\frac{n}{n-1}\right) =$$

$$2\mu(n-1)\exp\left(\ln\frac{n}{n-1}\right)\exp\left(-n\ln\frac{n}{n-1}\right) = 2n\mu\exp\left(-n\ln\frac{n}{n-1}\right)$$

Probability curve for all n markers matching:

p(n markers match) $= 2n\mu\exp(-2n\mu t)$

MRCA at k peak:

$t\left(\text{k peak}\right) = \frac{1}{2}\frac{\ln\frac{n}{k}}{\mu}$

MRCA at k=n-1 peak:

$t\left(\text{k=n-1 peak}\right) = \frac{1}{2}\frac{\ln\frac{n}{n-1}}{\mu}$ or $2\mu t = \ln\frac{n}{n-1}$

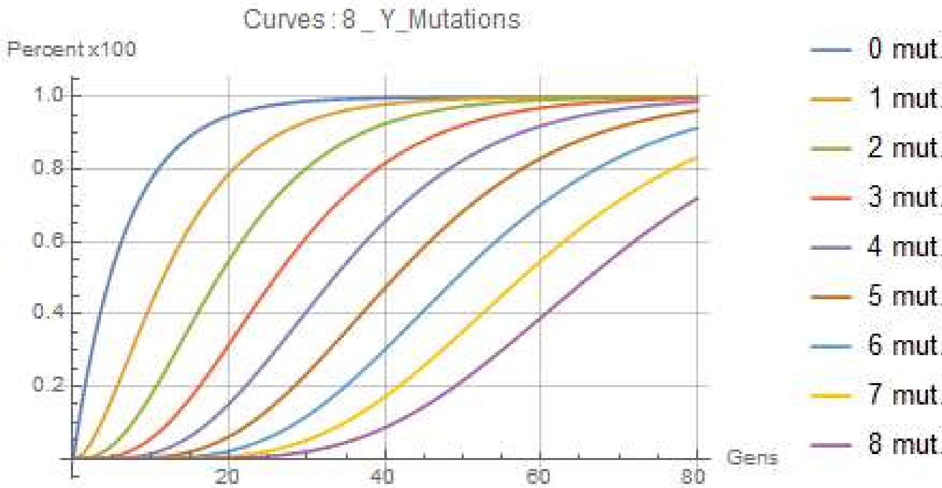
Therefore, probability curve for all n markers matching value at peak of k=n-1 curve:

p(n markers match at k=n-1 peak) $= 2n\mu\exp\left(-n\ln\frac{n}{n-1}\right)$

Therefore,

p(n markers match at k=n-1 peak) $= p$(k=n-1 peak) Q.E.D.

Integrating the curves above for 37 markers yields cumulative matching probability:



The vertical axis multiplied by 100 is the % probability. The horizonal axis is the generations (up t0 80) separated. The blue curve on the left is for 0 relative mutations and the yellow curve on the right is for 8 relative mutations.

Reference:
http://freepages.rootsweb.com/~craventaylors/genealogy/DNA/Probabilities-in-DNA-4.htm