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:

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

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:

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

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:

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

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=37:

- k=36: $t_p = 250\ln\frac{37}{36} = 6.8497$
- k=35: $t_p = 250\ln\frac{37}{35} = 13.892$
- k=34: $t_p = 250\ln\frac{37}{34} = 21.139$

n=59:

- k=58: $t_p = 250\ln\frac{59}{58} = 4.2736$
- k=57: $t_p = 250\ln\frac{59}{57} = 8.6215$
- k=56: $t_p = 250\ln\frac{59}{56} = 13.046$
- k=55: $t_p = 250\ln\frac{59}{55} = 17.551$
- k=54: $t_p = 250\ln\frac{59}{54} = 22.138$

Then the value at the peak is:

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

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

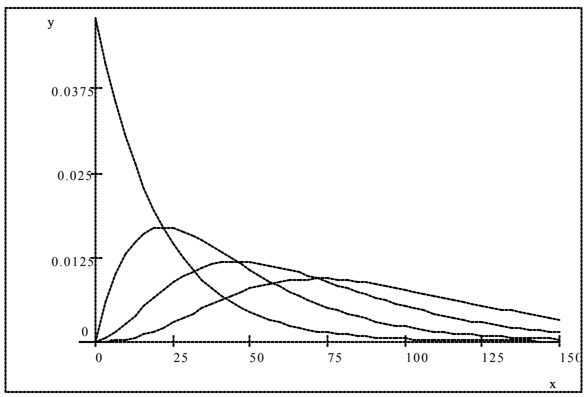
n=12; k=11, 10 & 9:

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

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

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

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

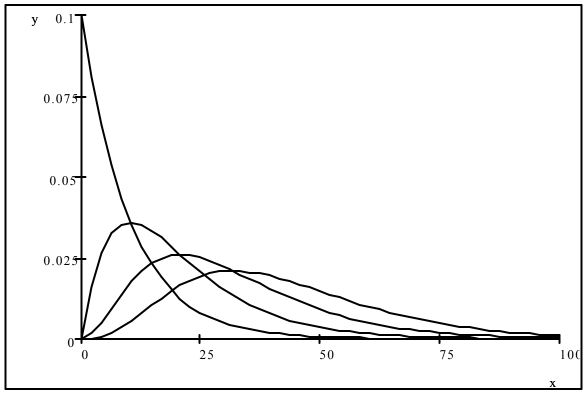


$$p_{25}\langle t\rangle = 2*25*\frac{1}{500}\exp(-2*25*\frac{1}{500}t) = \frac{1}{10}e^{-\frac{1}{10}t}$$

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

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

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

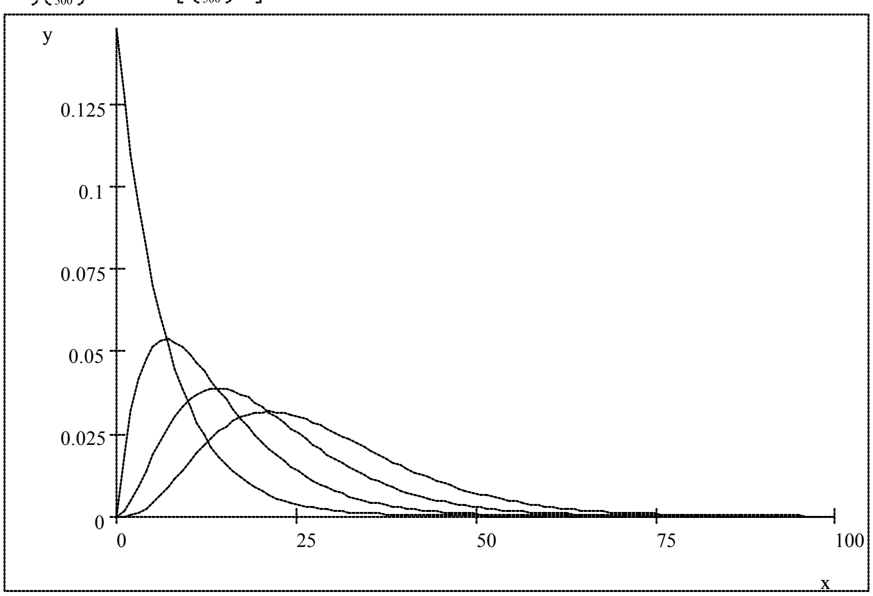


$$p_{37}\langle t\rangle = 2*37*\frac{1}{500}\exp\left(-2*37*\frac{1}{500}t\right) = 0.148e^{-0.148t}$$

$$p_{37,36}\langle t\rangle = \frac{\prod_{i=0}^{37-36}\left[2\left(\frac{1}{500}\right)(37-i)\right]}{2^{37-36}(37-36)!\left(\frac{1}{500}\right)^{37-36}}\frac{\left(1-\exp\left[-2\left(\frac{1}{500}\right)t\right]\right)^{37-36}}{\exp\left[2\left(\frac{1}{500}\right)36t\right]} : 5.328\frac{1.0-1.0e^{-0.004t}}{e^{0.144t}}$$

$$p_{37,35}\langle t\rangle = \frac{\prod_{i=0}^{37-35}\left[2\left(\frac{1}{500}\right)(37-i)\right]}{2^{37-35}(37-35)!\left(\frac{1}{500}\right)^{37-35}}\frac{\left(1-\exp\left[-2\left(\frac{1}{500}\right)t\right]\right)^{37-35}}{\exp\left[2\left(\frac{1}{500}\right)35t\right]} : 93.24\frac{\left(1.0-1.0e^{-0.004t}\right)^{2}}{e^{0.14t}}$$

$$p_{37,34}\langle t\rangle = \frac{\prod_{i=0}^{37-34}\left[2\left(\frac{1}{500}\right)(37-i)\right]}{2^{37-34}(37-34)!\left(\frac{1}{500}\right)^{37-34}}\frac{\left(1-\exp\left[-2\left(\frac{1}{500}\right)t\right]\right)^{37-34}}{\exp\left[2\left(\frac{1}{500}\right)34t\right]} : 1056.7\frac{\left(1.0-1.0e^{-0.004t}\right)^{3}}{e^{0.136t}}$$

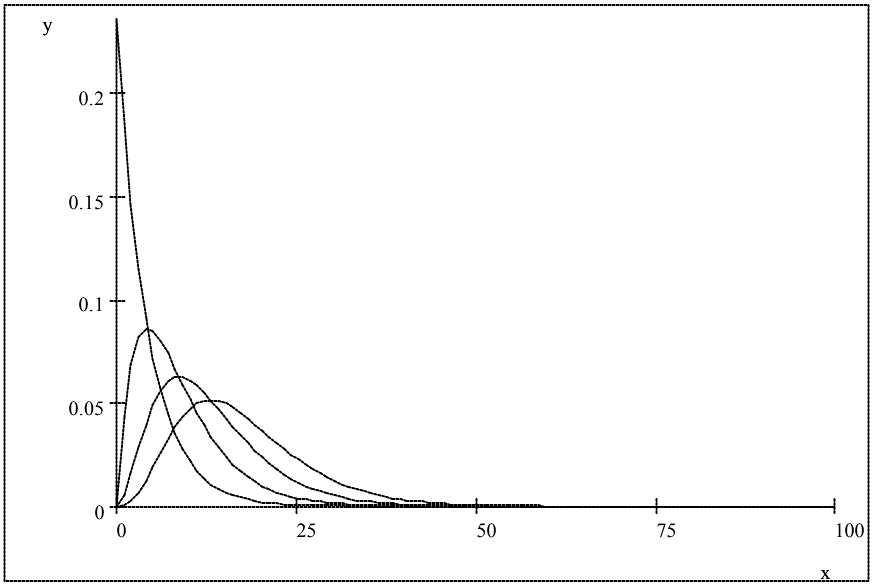


$$p_{59}\langle t\rangle = 2*59*\frac{1}{500}\exp\left(-2*59*\frac{1}{500}t\right) = 0.236e^{-0.236t}$$

$$p_{59,58}\langle t\rangle = \frac{\prod_{i=0}^{59-58}\left[2\left(\frac{1}{500}\right)(59-i)\right]}{2^{59-58}(59-58)!\left(\frac{1}{500}\right)^{59-58}}\frac{\left(1-\exp\left[-2\left(\frac{1}{500}\right)t\right]\right)^{59-58}}{\exp\left[2\left(\frac{1}{500}\right)58t\right]} : 13.688\frac{1.0-1.0e^{-0.004t}}{e^{0.232t}}$$

$$p_{59,57}\langle t\rangle = \frac{\prod_{i=0}^{59-57}\left[2\left(\frac{1}{500}\right)(59-i)\right]}{2^{59-57}(59-57)!\left(\frac{1}{500}\right)^{59-57}}\frac{\left(1-\exp\left[-2\left(\frac{1}{500}\right)t\right]\right)^{59-57}}{\exp\left[2\left(\frac{1}{500}\right)57t\right]} : 390.11\frac{\left(1.0-1.0e^{-0.004t}\right)^{2}}{e^{0.228t}}$$

$$p_{59,56}\langle t\rangle = \frac{\prod_{i=0}^{59-56}\left[2\left(\frac{1}{500}\right)(59-i)\right]}{2^{59-56}(59-56)!\left(\frac{1}{500}\right)^{59-56}}\frac{\left(1-\exp\left[-2\left(\frac{1}{500}\right)t\right]\right)^{59-56}}{\exp\left[2\left(\frac{1}{500}\right)56t\right]} : 7282.0\frac{\left(1.0-1.0e^{-0.004t}\right)^{3}}{e^{0.224t}}$$



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:

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

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

$t(\text{k-1 peak}) = \frac{1}{2}\frac{\ln\frac{n}{k-1}}{\mu}$ 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, $p(\text{k-1 peak}) = p(\text{k curve at k-1 peak})$ 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:

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

Probability curve for all n markers matching:

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

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:

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

Therefore,

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