We have two DNA sequences (say modern human and Neanderthal coding region mtDNA of length 15447 bases), denote them $X_1$ and $X_2$. They are separated by $T$ time (up and down the species tree in this case).

We assume mutations occur as a Poisson process in each site in the sequence. To analyze, we have a hierarchy of assumptions:

1. **Molecular clock**: Mutations occur at fixed rate across time in the evolution (possible violations: generation length, radioactivity...)

2. **Fixed rate across sites**: Mutations occur at same rate in all sites

3. **Binary genomes**: Genomes have two states

With all three assumptions, by comparing the two sequences we know that when $X_{1j} = X_{2j}$, site $j$ had even number of mutations, otherwise odd. For $Y \sim \text{Pois}(\lambda T)$, the probability of being even is $P(Y \text{even}) = 0.5 + 0.5 \exp(-2\lambda T)$.

Hence if the sequences differ at $M$ loci, we have $M \sim \text{Bin}(15447, 0.5 - 0.5 \exp(-2\lambda T))$, and then we can find MLE of $\lambda$ given $T$ and vice versa from $\hat{p} = M/15447$ by invariance of MLE:

$$\hat{\lambda} = \frac{-\log(1 - 2\hat{p})}{2T}, \quad \hat{T} = \frac{-\log(1 - 2\hat{p})}{2\hat{\lambda}}.$$

Now if we want to estimate $T(\text{modern, Neanderthal})$ we can note that:

1. Modern humans and Neanderthals differ at 170 sites, so $\hat{p}_1 = 0.011$.

2. Modern humans and Chimpanzees differ at 1300 sites, so $\hat{p}_2 = 0.084$.

3. We also assume that for humans and Chimpanzees, $T_2 = 13M$ years (6.5MY to last common ancestor).

So calibration gives us $\hat{\lambda} = \frac{-\log(1 - 2 \times 0.084)}{2T_2} = 7.1 \times 10^{-9}$, and for the entire mtDNA coding region: $\lambda_{\text{tot}} = 15447 \times 7.1 \times 10^{-9} = 1.09 \times 10^{-4}$.

Using this, we can estimate $T_1$: $\frac{-\log(1 - 2 \times 0.011)}{2\hat{\lambda}} = 1.57 \times 10^6$.

Conclusion: $1.57 \times 10^6 / 2 = 785K$ years since the split.

**Note:** A simplified calculation assuming that the number of mutations is 0 or 1, gives $\hat{\lambda} = \hat{p}/13M = 6.5 \times 10^{-9}$, not much different, because $\hat{p} << 1$. 
1 Relaxing Assumption 2

Assumption 1 is quite inevitable, and 3 is not major because transitions are much more common than transversions.

We would like to test Assumption 2 statistically. Denote the number of mutations in site \( i \) in \( Z_i \sim Pois(\lambda_i T) \). As an alternative, we may take the Negative Binomial distribution:

\[
H_0 : \ 
\lambda_1 = \ldots = \lambda_{15447} \\
H_A : \ Z_i \sim NB(\alpha, p) : P(Z_i = k) = \frac{\Gamma(k + \alpha)}{\Gamma(k + 1)\Gamma(\alpha)} (1 - p)^\alpha p^k.
\]

This is the famous approach of Tamura and Nei (1992).

**Reminder:** Gamma distribution: \( X \sim \Gamma(\alpha, \beta) \) has density \( f(x) = \frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)} \). If \( \lambda \sim \Gamma(\alpha, \beta) \) and \( Z|\lambda \sim Pois(\lambda) \), then unconditionally \( Z \sim NB(\alpha, p = 1/(1 + \beta)) \). This is often called "Overdispersed Poisson" and can be thought of as a random effects model. **Proof:**

\[
P(X = k) = \int_0^\infty \Gamma(\lambda) \cdot p_{Pois}(\lambda)(k)d\lambda = \int_0^\infty \lambda^{\alpha-1}e^{-\lambda\beta} \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\lambda} \frac{\lambda^k}{\Gamma(k + 1)} d\lambda = \\
= \left( \frac{\beta}{\beta + 1} \right)^\alpha \left( \frac{1}{\beta + 1} \right)^k \frac{\Gamma(k + \alpha)}{\Gamma(k + 1)\Gamma(\alpha)} \int_0^\infty \lambda^{k + \alpha - 1}e^{-\lambda(\beta + 1)} \frac{1}{\Gamma(k + \alpha)} d\lambda = \\
= \left( \frac{\beta}{\beta + 1} \right)^\alpha \left( \frac{1}{\beta + 1} \right)^k \frac{\Gamma(k + \alpha)}{\Gamma(k + 1)\Gamma(\alpha)} = NB(k; \alpha, p = 1/(1 + \beta)).
\]

**Notes:**

1. Moments of NB: \( E(X) = E(E(X|\lambda)) = E(\Gamma(\alpha, \beta)) = \alpha/\beta \) (iterated expectation). Variance (Law of total variation):

\[
Var(X) = Var(E(X|\lambda)) + E(Var(X|\lambda)) = Var(\lambda) + E(\lambda) = \frac{\alpha}{\beta^2} + \frac{\alpha}{\beta} = \frac{p\alpha}{(1 - p)^2}.
\]

2. As \( \alpha \to \infty \) with \( \alpha/\beta = \lambda \) fixed, we converge to Poisson (the \( \Gamma \) gets peaked at the point \( \alpha/\beta \)).

**Data analysis and hypothesis test:** see paper and code on class page, we get \( \hat{\alpha} = 0.168 \).

**Probability of \( Z \sim NB(\alpha, \beta) \) to be even:**

\[
P(\text{Even}) = \int_0^\infty \Gamma(\lambda) \cdot P(Pois(\lambda)\text{even})d\lambda = \int_0^\infty \lambda^{\alpha-1}e^{-\lambda\beta} \frac{\beta^\alpha}{\Gamma(\alpha)} (0.5 + 0.5e^{-2\lambda}) d\lambda = \\
= \left( 0.5 + 0.5 \left( \frac{\beta}{\beta + 2} \right)^\alpha \right) \int_0^\infty \lambda^{\alpha-1}e^{-\lambda(\beta + 2)} \frac{1}{\Gamma(\alpha)} d\lambda = 0.5 + 0.5 \left( \frac{\beta}{\beta + 2} \right)^\alpha
\]
Calculation of split time using $NB(0.168, \cdot)$: Chimpanzee: $0.5 + 0.5(\beta/(\beta + 2))^{0.168} = 0.916 \Rightarrow \hat{\beta} = 1.003$. This gives average of $\frac{\hat{\alpha}}{\beta \cdot 13 \cdot 10^6} = 1.28 \times 10^{-8}$ mutations per site. Hence $\lambda_{tot} = 1/5033.5$. About half of what we had with Poisson calculation!

Neanderthal: $0.5 + 0.5(\beta/(\beta + 2))^{0.168} = 0.989 \Rightarrow \hat{\beta} = 14.1$. Using the calibration this gives $T_{Nean} = 0.5 \times \frac{\hat{\alpha}}{\beta \cdot 1.28 \cdot 10^{-8}} = 496K$.

Conclusion: 496$K$ years from modern-Neanderthal split using NB, compared to almost double using Poisson.