

Notes on the Poisson Process and Models of Molecular Evolution

Bret Larget

January 23, 2009

A Simple Model of Molecular Evolution

We begin the discussion with a question: How can we use observed DNA sequence data from humans and chimpanzees to make statistical inferences about their time of divergence?

The data we have at hand are an alignment of two 235 base-pair DNA sequences. The data comes from the hypervariable I region of the mitochondrial genome. The mitochondrial genome in humans and chimps is a ring of DNA of about 16,500 base pairs that is passed maternally to offspring and is separate from the 3 billion+ nucleotides in the chromosomes in the cell nuclei. In a DNA sequence, each base is a letter from the alphabet $\{A, C, G, T\}$ and each letter stands for a base made up of several dozen atoms. A DNA sequence can be thought of as a word in this small alphabet.

Here is the data at the first fifteen sites.

```
human  CCAAGTATTGACTCA...
chimp  CTAAGTATTGGCTTA...
```

Note that the bases are the same for 12 of these sites and differ for 3. In the full data set, there are 39 differences among the 235 sites.

The main idea underlying evolution is that humans and chimps share a common ancestor. Here, if we traced back the maternal lineage from the specific human and chimp sampled, we would eventually (after several million years) find the individual in the human/chimp ancestral species that is the most recent common ancestor of the two individuals sampled in present times following only maternal lineages.

The statistical idea is to develop a model for how DNA changes over time and in the context of this model to estimate the divergence times of human and chimpanzees. Any such model will incorporate several assumptions. The model we describe here will include these assumptions:

1. The common ancestor A had a 235 base-pair sequence from which both current sequences descended.
2. All 235 sites evolve independently, so it suffices to describe only the model for a single site.
3. Given the base at the ancestor A , the bases at the human H and chimp C are conditionally independent.
4. Nucleotide substitutions along the two lineages from A to H and from A to C behave as Poisson processes with rate λ .
5. At each site, the value of the base at A is uniformly chosen among the four possible bases.
6. At each substitution even, all three possible substitutions are equally likely.

The third assumption can be written as

$$\mathbb{P}(H_i = x, C_i = y \mid A_i = z) = \mathbb{P}(H_i = x \mid A_i = z) = \mathbb{P}(C_i = y \mid A_i = z)$$

where A_i , H_i , and C_i are the bases at the i th site for the ancestor, human, and chimp. With these assumptions, the probability distribution of the observed data can be written as

$$\begin{aligned} & \mathbb{P}(H = (x_1, \dots, x_n) \cap C = (y_1, \dots, y_n)) \\ &= \prod_{i=1}^n \left(\sum_k \mathbb{P}(A_i = k) \mathbb{P}(H_i = x_i \mid A_i = k) \mathbb{P}(C_i = y_i \mid A_i = k) \right) \end{aligned}$$

where the transition probabilities are determined by the Poisson process. We need to describe more fully the Poisson process in order to better understand the model.

Poisson Process

There are multiple equivalent ways to specify a Poisson process, which (in one dimension) can be thought of as a way to place random points on a line. Typically the line is thought of as *time* and the points are *events*.

Waiting time description.

Let W_1, W_2, \dots be independent and identically distributed (i.i.d) Exponential(λ) random variables (with density $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$). Define $S_0 = 0$ and let $S_n = \sum_{i=1}^n W_i$ be the sum of the first n W_i . Note that S_n is the time of the n th event and that W_i is the length of the i th interevent time.

Define $N_t = \max\{n : S_n \leq t\}$ for all $t \geq 0$. Note that N_t counts the number of events in the time up to t and that for $0 < s < t$, $N_t - N_s$ counts the number of events after time s up to time t .

Under these assumptions, the following facts follow.

1. $N_t \sim \text{Poisson}(\lambda t)$ for all $t > 0$.
2. For $0 \leq s < t$, $N_t - N_s \sim \text{Poisson}(\lambda(t - s))$ for all $t > 0$.
3. For $0 = t_0 < t_1 < t_2 < \dots < t_k$, the random variables $X_i = N_{t_i} - N_{t_{i-1}}$ are mutually independent for $i = 1, \dots, k$.

So i.i.d. exponential waiting times lead to Poisson counts in fixed intervals.

Poisson description.

Let $t > 0$ be fixed and let $X \sim \text{Poisson}(\lambda)$. If $X = 0$, then there are no events in the interval $[0, t]$. If $X = n > 0$, then let $U_1, \dots, U_n \sim \text{i.i.d. Uniform}(0, t)$ and let $U_{(1)} < U_{(2)} < \dots < U_{(n)}$ be the order statistics. The joint probability distribution of these points is identical to that from the exponential waiting time description. Specifically, given $X = n$, the random variables $Y_i = U_{(i)} - U_{(i-1)}$ for $i = 1, \dots, n$ (letting $U_{(0)} = 0$) have the same conditional distribution as W_1, \dots, W_n given that $S_n \leq t$ and $S_{n+1} > t$.

General Description

The Poisson process turns out to be the unique stochastic process that arises from three relatively general assumptions: homogeneity, independence, and no simultaneous events. Specifically, if a general point process that places points on the interval $(0, t)$ where N_s counts the points in $(0, s)$ satisfies these assumptions:

1. **homogeneity:** for all $0 \leq u < v \leq t$, $\mathbf{E}(N_v - N_u) = \lambda(v - u)$.
2. **independence:** for $0 = t_0 < t_1 < \dots < t_n = t$, the random variables $N_{t_i} - N_{t_{i-1}}$ are independent.
3. **no simultaneous events:** for all $u, s > 0$, $\mathbf{P}(N_{u+s} - N_s = 1) = \lambda u + o(u)$ and $\mathbf{P}(N_{u+s} - N_s > 1) = o(u)$

then, it is a Poisson process equivalent to the previous two descriptions. The last assumption can be stated equivalently as

$$\lim_{u \downarrow 0} \frac{\mathbf{P}(N_{u+s} - N_s = 1)}{u} = \lambda$$

and that

$$\lim_{u \downarrow 0} \frac{\mathbf{P}(N_{u+s} - N_s > 1)}{u} = 0.$$

In words, the probability of two or more events in small time intervals gets closer to zero faster than the rate that the interval goes to zero.

A Likelihood Model

It follows from some further advanced mathematics that under a Poisson process, the transition probabilities above take the form

$$P(H_i = z | A_i = k) = P(C_i = z | A_i = k) = \begin{cases} 1/4 + (3/4)e^{-(4/3)\lambda t} & \text{if } z = k \\ 1/4 - (1/4)e^{-(4/3)\lambda t} & \text{if } z \neq k \end{cases}.$$

The probability of a specific outcome where the sites agree such as AA is then

$$\begin{aligned} P(H_i = x, C_i = x) &= \sum_k P(A_i = k) P(H_i = x | A_i = k) P(C_i = x | A_i = k) \\ &= P(A_i = x) P(H_i = x | A_i = x) P(C_i = x | A_i = x) \\ &\quad + \sum_{k \neq x} P(A_i = k) P(H_i = x | A_i = k) P(C_i = x | A_i = k) \\ &= (1/4) \left(1/4 + (3/4)e^{-(4/3)\lambda t}\right)^2 + 3 \times (1/4) \left(1/4 - (1/4)e^{-(4/3)\lambda t}\right)^2 \\ &= \left(1/4 + (3/4)e^{-(8/3)\lambda t}\right) / 4. \end{aligned}$$

The last inequality follows algebraically, but it also is the expression from a Poisson process with rate λ over time $2t$ that begins and ends at x .

The probability of a specific outcome where the sites disagree such as AC is then for $x \neq y$

$$\begin{aligned} P(H_i = x, C_i = y) &= \sum_k P(A_i = k) P(H_i = x | A_i = k) P(C_i = y | A_i = k) \\ &= P(A_i = x) P(H_i = x | A_i = x) P(C_i = y | A_i = x) \\ &\quad + P(A_i = y) P(H_i = x | A_i = y) P(C_i = y | A_i = y) \\ &\quad + \sum_{k \neq x, y} P(A_i = k) P(H_i = x | A_i = k) P(C_i = x | A_i = k) \\ &= 2 \times (1/4) \left(1/4 + (3/4)e^{-(4/3)\lambda t}\right) \left(1/4 - (1/4)e^{-(4/3)\lambda t}\right) + 2 \times (1/4) \left(1/4 - (1/4)e^{-(4/3)\lambda t}\right)^2 \\ &= \left(1/4 - (1/4)e^{-(8/3)\lambda t}\right) / 4. \end{aligned}$$

Again, the last inequality follows algebraically, but it also is the expression from a Poisson process with rate λ over time $2t$ for the probability of going from x to y .

This can all be put together to write an expression for the probability of each possible pair of human and chimp sequences that agree at x sites and disagree at $n - x$ sites.

$$P(x | \lambda, t) = \left(\left(1/4 - (1/4)e^{-(8/3)\lambda t}\right) / 4 \right)^x \left(\left(1/4 + (3/4)e^{-(8/3)\lambda t}\right) / 4 \right)^{n-x}$$

which can be rewritten as

$$P(x | \lambda, \theta) = (1/4)^{4n} \theta^x (4 - 3\theta)^{n-x}$$

with some rearrangement and letting $\theta = 1 - e^{-(8/3)\lambda t}$ (and then $1 + 3e^{-(8/3)\lambda t} = 4 - 3\theta$) so that $t = -3 \log(1 - \theta) / (8\lambda)$.

Estimation

This derivation provides a framework for estimating t from real data. We can first use the simpler expression involving θ to estimate it from real data and then transform to get t . Many studies have been done to examine the substitution rate per in HV1 in humans. While there is considerable discussion about this, we will use $\lambda = 0.025$ substitutions per million years as an estimate.

One way to estimate θ from the real data is to use the principle of *maximum likelihood*. Under this principle, the best point estimate for θ is that parameter value that maximizes the probability of the observed data. So, the

likelihood function is the same as the probability function of the data, but the roles of what is fixed and what varies change. We can write the likelihood as

$$L(\theta) = (1/4)^{4n} \theta^x (4 - 3\theta)^{n-x} .$$

It is almost always the case that the maximization problem is greatly simplified by first taking logarithms. If we let $\ell(\theta) = \log L(\theta)$, it follows that

$$\ell(\theta) = -4n \log(4) + x \log \theta + (n - x) \log(4 - 3\theta)$$

and the derivative is

$$\ell'(\theta) = \frac{x}{\theta} - \frac{3(n-x)}{4-3\theta} = 0$$

which is solved when

$$\hat{\theta} = \frac{4x}{3n} .$$

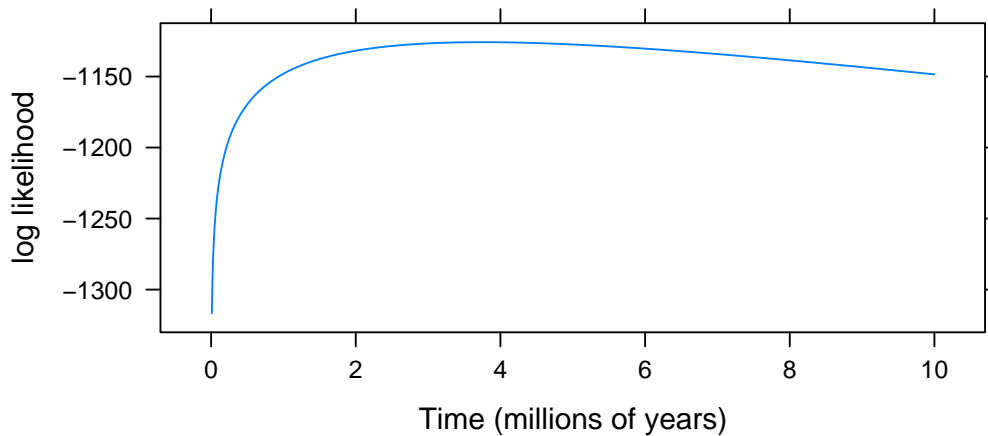
This is the maximum likelihood estimate of θ . But we need to transform this for our primary question of interest. The maximum likelihood estimate of t is then

$$\hat{t} = -3 \log\left(1 - \frac{4x}{3n}\right) / (8\lambda) = -3 \log\left(1 - \frac{4(39)}{3(235)}\right) / (8(0.025)) \doteq 3.75 ,$$

million years ago.

A graph of the log likelihood versus t can be used to check that the derivations were correct.

```
> library(lattice)
> u = seq(0, 10, 0.01)
> lambda = 0.025
> theta = 1 - exp(-(8/3) * lambda * u)
> x = 39
> n = 235
> logLikelihood = -4 * n * log(4) + x * log(theta) + (n - x) *
+   log(4 - 3 * theta)
> print(xyplot(logLikelihood ~ u, type = "l", xlab = "Time (millions of years)",
+   ylab = "log likelihood"))
```



Quantifying Uncertainty

Statistical point estimates of parameter values without an estimate of accuracy are generally not sufficient. In some simple cases, there are easy formulas for standard errors. In other cases, we can use computational methods.

One computational method is the *parametric bootstrap*. The underlying idea is deep, but the process is intuitive. For the given problem, we have an estimate. The key idea is to create new data sets where the estimated model is true and apply the estimation procedure to the new data. Variation of the simulated data around the estimate should be similar to variation of the estimate around the true parameter value. We can, for example, generate 10,000 new data sets from the estimated model, estimate t for each, and find the middle 95% of the estimates from simulated data as an approximate 95% confidence interval for the original estimate from the data.

Here is some R code to accomplish this. Notice that the probability of observing a site of data that agrees is the sum of four probabilities

$$4 \times (1/4)(1/4 + (3/4)e^{-(8/3)(\lambda * t)}) = (1 + 3e^{-(8/3)(\lambda * t)})/4$$

and that the probability of observing data that does not agree is the sum of 12 probabilities,

$$12 \times (1/4)(1/4 - (1/4)e^{-(8/3)(\lambda * t)}) = 3(1 - e^{-(8/3)(\lambda * t)})/4$$

We can thus generate x from the fitted model using the binomial distribution.

```
> x = 39
> n = 235
> lambda = 0.025
> set.seed(353)
> estTime = function(x, n, lambda) {
+   return(-3/8/lambda * log(1 - 4 * x/3/n))
+ }
> t.hat = estTime(x, n, lambda)
> sim.p = 3 * (1 - exp(-8 * lambda * t.hat/3))/4
> sim.x = rbinom(1e+05, n, sim.p)
> sim.t = estTime(sim.x, n, lambda)
> quantile(sim.t, c(0.025, 0.975))

      2.5%      97.5%
2.595051 5.123876
```

So, based on the data and model, we are 95% confident that the diversion time between humans and chimps is between 2.6 and 5.1 million years ago.

Model Checking

This estimated divergence time is rather uncertain, but is also substantially lower than those made by many experts. What might have gone wrong?

The model we used makes a number of assumptions that are violated, some badly. For example, we assume that all possible bases are equally likely and that all substitutions are equally likely. This is inconsistent with the observed data. We can summarize the observed data by constructing a table.

```
> library(ape)
> hc = read.dna("human-chimp.txt", format = "fasta", as.character = T)
> x = sum(hc[[1]] != hc[[2]])
> n = length(hc[[1]])
> print(x)

[1] 39

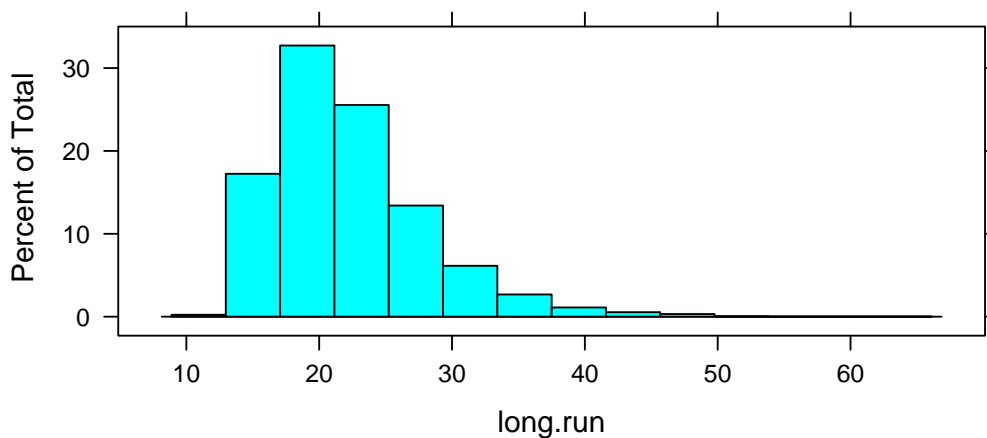
> print(n)
```


[1] 51

We can now repeat this for 10,000 random permutations of the data and graph this distribution.

```
> long.run = rep(NA, 10000)
> for (i in 1:10000) {
+   long.run[i] = longestRun(sample(y))
+ }
> print(histogram(long.run))
> print(quantile(long.run, c(0.025, 0.975)))
```

```
2.5% 97.5%
 14   37
```



Notice how unusually long the observed length of 51 is compared to permutations of the data. This is consistent with a story of *hot spots* of high rate variation.

Conclusion

A complete statistical analysis often involves numerical and graphical data summaries, stochastic modeling, statistical inference, and model checking and refinement. For this analysis, the model checking indicates several features in the data ignored by the model. It is unclear what effect this might have on the final analysis, but it is evident that this would need to be examined more closely before trusting the result of the analysis.