Statistical Genetics, Spring 2024

Class notes 1, part 2

Estimation with two sequences

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...), also included is assumption of independence between sites
2. **Fixed rate across sites:** Mutations occur at same rate in all sites
3. **Binary genomes:** Genomes have two states (this assumption is pretty mild because transitions are a lot more common than transversions)
4. **Known mutation count in each site:** This is a (naive) assumption that in each site we know not only the states of the two sequences, but how many mutations occurred between them

With all four assumptions, if we have known how many mutation occurred in each site, denote them $Y_1, \ldots, Y_{15447}$, we would know that $Y_i \sim Pois(\lambda T)$, and from Poisson properties the summary statistic is the sum of mutations with distribution:

$$S = \sum_{i=1}^{15447} Y_i \sim Pois(\lambda_{tot} T) \ , \ \lambda_{tot} = 15447\lambda$$

and corresponding trivial MLEs: given T: $\hat{\lambda_{tot}} = S/T$ and given λ_{tot}: $\hat{T} = S/\lambda_{tot}$.
It is important to note that given Assumption 4, the same holds even without Assumption 2 (since the sum of independent Poissons is Poisson even with different rates), with λ_{tot} the parameter of interest.

But Assumption 4 that we know the Y_i's is naive and unrealistic if we only have the sequences, so we dispense with it completely. Rather with the other three assumptions, by comparing the two sequences we know that when $X_{1j} = X_{2j}$, site j had even number of mutations, otherwise odd.

Claim: For $Y \sim Pois(\lambda T)$, the probability of being even is $P(Yeven) = 0.5 + 0.5\exp(-2\lambda T)$.

Proof: Assume we have two independent $Pois(\lambda T)$ random variables. The probability they both have zero mutations is $\exp(-2\lambda T)$. Now take the last event that happened on either Poisson. With probability $1/2$ it happened on the first and flipped its parity, with $1/2$ in the second and the first's parity is unchanged. Hence conditional on $Y_1 + Y_2 > 0$ the probability of Y_1 being even and odd is equal:

$$\mathbb{P}(Y_1 even) = \exp(-2\lambda T) + 0.5 \times (1 - \exp(-2\lambda T)) = 0.5 + 0.5\exp(-2\lambda T).$$

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

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

(Note this is not an unbiased estimate).

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

1. Modern humans and Neanderthals differ at around 170 sites, so $\hat{p}_1 = 0.011$.
2. Modern humans and Chimpanzees differ at around 1300 sites, so $\hat{p}_2 = 0.084$.
3. We also assume that for humans and Chimpanzees, a common estimate based on fossils is $T_2 = 13M$ years ($6.5MY$ to last common ancestor).

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

Using this, we can estimate T_1: $\frac{-\log(1-2\cdot 0.011)}{2\hat{\lambda}} = 1.57 \times 10^6$.
Conclusion: $1.57 \times 10^6/2 = 785K$ years since the split.

Note: A simplified calculation using Assumption 4 and 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 $\hat{p} << 1$.

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 by $Z_i \sim Pois(\lambda_i T)$. As an alternative, we may take the Negative Binomial distribution:

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

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

Reminder: Gamma distribution: $X \sim \Gamma(\alpha, \beta)$ has density $f_\Gamma(x) = x^{\alpha-1}e^{-x\beta}\frac{\beta^\alpha}{\Gamma(\alpha)}$. 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:

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

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 Γ gets peaked at the point α/β).

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:

$$\begin{aligned} P(Zeven) &= \int_0^\infty f_\Gamma(\lambda) \cdot P(Pois(\lambda)even)d\lambda = \int_0^\infty \lambda^{\alpha-1}e^{-\lambda\beta}\frac{\beta^\alpha}{\Gamma(\alpha)}(0.5 + 0.5e^{-2\lambda})d\lambda = \\ &= 0.5 + \left(0.5\left(\frac{\beta}{\beta+2}\right)^\alpha\right) \int_0^\infty \lambda^{\alpha-1}e^{-\lambda(\beta+2)}\frac{(\beta+2)^\alpha}{\Gamma(\alpha)}d\lambda = 0.5 + 0.5\left(\frac{\beta}{\beta+2}\right)^\alpha \end{aligned}$$

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 $\frac{\hat{\alpha}}{\hat{\beta}\cdot 13\cdot 10^6} = 1.28 \times 10^{-8}$ 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 $T_{nean} = 0.5 \times \frac{\hat{\alpha}}{\hat{\beta}\cdot 1.28\cdot 10^{-8}} = 496K$.

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

Summary and conclusions

1. The assumptions we take can have a major effect on genetic estimates
2. In this case, with the (clearly incorrect) assumption of fixed rate we get an estimate that is about twice as big as the one without the fixed rate assumption
3. We have a lot more to do: relax the binary assumption, critically examine the molecular clock assumption and violations. But these are expected to have a minor effect on estimates compared to the fixed rate assumption
4. Our analysis is still a bit naive in several aspects. For example, we estimate α based on the modern human tree, then use the estimate as the true parameter in estimating branch length. It makes more sense and is a more correct statistical approach to do all estimation simultaneously as part of one big model. But it makes the analysis substantially more complicated

For more, see:
Levinstein-Hallak and Rosset (2024). Dating ancient splits in phylogenetic trees, with application to the human-Neanderthal split