Lecture 23 : Estimating the recombination rate

MATH285K - Spring 2010

Lecturer: Sebastien Roch

References: [Dur08, Chapter 3.2].

Previous class

Recall that for a two-locus recombination process without mutation (the loci are called a and b):

THM 23.1 (Tree-Length Covariance: Recursion) *Let $x = (i, j, k)$ be the initial state where i (respectively j, and k) is the number of lineages with only a (respectively only b, and both a and b) material ancestral to the samples with $n_a = i + k$, $n_b = j + k$, and $\ell = i + j + k$. Let $F(x)$ be the covariance of the tree lengths τ_a and τ_b started at x. If X is the state after the first jump. Then*

$$F(x) = \mathbb{E}_x[F(X)] + \frac{2k(k-1)}{\beta_x(n_a - 1)(n_b - 1)},$$

where

$$\beta_x = \frac{\ell(\ell-1) + k\rho}{2},$$

and $\rho/2$ is the recombination rate per lineage.

An application of this theorem to the 2-sample case gives:

THM 23.2 (Covariance: Two-Sample Case) *We have*

$$F(0,0,2) = 4\frac{\rho + 18}{\rho^2 + 13\rho + 18},$$

$$F(1,1,1) = 4\frac{6}{\rho^2 + 13\rho + 18},$$

and

$$F(2,2,0) = 4\frac{4}{\rho^2 + 13\rho + 18}.$$

(The factor of 4 comes from the difference between coalescence time and tree length.)

1 Mutation model

It is not entirely obvious to extend the infinite-sites model to the case with recombination. Indeed, the linear order of the sites is now important. One way to deal with this is to arrange m infinite-sites loci linearly with mutation rates $\frac{\theta}{2m}$ and recombination rate $\frac{\rho}{2(m-1)}$ between any two consecutive loci. *There is no intra-locus recombination.* We then take a limit $m \to +\infty$.

Our goal in this lecture is to estimate the recombination rate. To do so, we must also estimate the mutation rate. We describe an approach based on pairwise differences. Let

$$\Delta_n \equiv \sum_{a=1}^{m} \Delta_n^a \equiv \frac{1}{\binom{n}{2}} \sum_{\{i,j\}} \Delta_{i,j} \equiv \sum_{a=1}^{m} \frac{1}{\binom{n}{2}} \sum_{\{i,j\}} \Delta_{i,j}^a,$$

where $\Delta_{i,j}^a$ is the number of differences between sequences i and j at locus a. Recall that

$$\mathbb{E}[\Delta_n] = m\mathbb{E}\left[\Delta_n^1\right] = m\left(\frac{\theta}{m}\right) = \theta.$$

(Recall also that (as proved in [Dur08])

$$\mathrm{Var}\left[\Delta_n^1\right] = \left(\frac{\theta}{m}\right)\frac{n+1}{3(n-1)} + \left(\frac{\theta}{m}\right)^2 \frac{2(n^2+n+3)}{9n(n-1)}.) \tag{1}$$

So $\theta_\pi = \Delta_n$ provides an estimate of θ. To estimate ρ, we need a quantity involving correlations between sites. A natural idea is to consider the sample variance of the pairwise differences, that is,

$$S_\pi^2 = \frac{1}{\binom{n}{2}} \sum_{\{i,j\}} (\Delta_{i,j} - \Delta_n)^2.$$

We will prove the following:

THM 23.3 *In the limit $m \to \infty$*

$$\mathbb{E}[S_\pi^2] = \theta\frac{2(n-2)}{3(n-1)} + \theta^2 g(\rho, n),$$

where g is a function given in [Dur08].

Recall that θ_π is not a consistent estimator of θ. Hence, to estimate θ^2 we use a corrected version θ_π^2. This will follow from:

THM 23.4 *In the limit $m \to \infty$,*

$$\mathrm{Var}[\Delta_n] = \theta \frac{n+1}{3(n-1)} + \theta^2 f(\rho, n),$$

where

$$f(\rho, n) = \frac{1}{\binom{n}{2}} \int_0^1 2(1-x) \frac{(\rho x) + (2n^2 + 2n + 6)}{(\rho x)^2 + 13(\rho x) + 18} \mathrm{d}x,$$

can be computed explicitly (see [Dur08]).

In particular, note that

$$\mathbb{E}[\theta_\pi^2] = \mathrm{Var}[\theta_\pi] + (\mathbb{E}[\theta_\pi])^2 = \theta \frac{n+1}{3(n-1)} + \theta^2 (f(\rho, n) + 1).$$

Hence, an unbiased estimator of θ^2 is

$$\gamma_\pi(\rho) = \frac{\theta_\pi^2 - [(n+1)/(3(n-1))]\theta_\pi}{f(\rho, n) + 1}.$$

Putting all this together, an estimate of ρ is given by the solution of

$$S_\pi^2 = \theta_\pi \frac{2(n-2)}{3(n-1)} + \gamma_\pi(\rho) g(\rho, n).$$

2 Proofs

We prove the two previous theorems. We begin with the second one.
Proof:(Theorem 23.4) Expanding the variance of Δ_n, the first term gives the term not depending on ρ

$$\mathrm{Var}[\Delta_n] = \sum_{a=1}^m \mathrm{Var}\left[\frac{1}{\binom{n}{2}} \sum_{\{i,j\}} \Delta_{i,j}^a\right] + \sum_{a \neq b} \mathrm{Cov}\left[\frac{1}{\binom{n}{2}} \sum_{\{i,j\}} \Delta_{i,j}^a, \frac{1}{\binom{n}{2}} \sum_{\{k,\ell\}} \Delta_{k,\ell}^b\right],$$

and

$$\sum_{a=1}^m \mathrm{Var}\left[\frac{1}{\binom{n}{2}} \sum_{\{i,j\}} \Delta_{i,j}^a\right] \to \theta \frac{n+1}{3(n-1)},$$

as $m \to \infty$, where we used (1). Rewriting the second term as

$$\sum_{a\neq b} \mathrm{Cov}\left[\frac{1}{\binom{n}{2}}\sum_{\{i,j\}}\Delta^a_{i,j}, \frac{1}{\binom{n}{2}}\sum_{\{k,\ell\}}\Delta^b_{k,\ell}\right] = \frac{1}{\binom{n}{2}^2}\sum_{a\neq b}\sum_{\{i,j\}}\sum_{\{k,\ell\}}\mathrm{Cov}\left[\Delta^a_{i,j}, \Delta^b_{k,\ell}\right],$$

we need to compute $\mathrm{Cov}\left[\Delta^a_{i,j}, \Delta^b_{k,\ell}\right]$. By conditioning on the tree lengths $\tau^a_{i,j}$ of locus a between i and j and $\tau^b_{k,\ell}$, we get

$$\mathrm{Cov}\left[\Delta^a_{i,j}, \Delta^b_{k,\ell}\right] = \left(\frac{\theta}{2m}\right)^2 \mathrm{Cov}\left[\tau^a_{i,j}, \tau^b_{k,\ell}\right].$$

Let

$$z = |b-a|\frac{\rho}{m-1},$$

be the total recombination rate between loci a and b. Then, using an argument similar to the one we used to compute the variance of the homozygosity,

$$\begin{aligned}\sum_{\{i,j\}}\sum_{\{k,\ell\}}&\mathrm{Cov}\left[\Delta^a_{i,j}, \Delta^b_{k,\ell}\right]\\ &= \left(\frac{\theta}{2m}\right)^2 \frac{4\binom{n}{2}}{z^2+13z+18}\left[(z+18)\cdot 1 + 6\cdot 2(n-2) + 4\cdot\binom{n-2}{2}\right]\\ &= \left(\frac{\theta}{m}\right)^2\binom{n}{2}\frac{z+(2n^2+2n+6)}{z^2+13z+18}.\end{aligned}$$

Summing over all values of $h = |b-a|$ and noting that there are $2(m-h)$ possibilities for each,

$$\begin{aligned}\frac{1}{\binom{n}{2}^2}&\sum_{a\neq b}\sum_{\{i,j\}}\sum_{\{k,\ell\}}\mathrm{Cov}\left[\Delta^a_{i,j}, \Delta^b_{k,\ell}\right]\\ &= \theta^2\frac{1}{\binom{n}{2}}\sum_{h=1}^{m}\frac{1}{m}\frac{2(m-k)}{m}\frac{\frac{\rho h}{m-1}+(2n^2+2n+6)}{(\frac{\rho h}{m-1})^2+13\frac{\rho h}{m-1}+18}.\end{aligned}$$

Taking a limit $m \to \infty$ and using a Riemann integral approximation gives the result. To compute the integral, factor the denominator. ■

We can now prove the first theorem.

Proof:(Theorem 23.3) This calculation is rather straightforward (up to a "miracle"; see [Dur08]). Rewrite

$$S^2_\pi = \left[\frac{1}{\binom{n}{2}}\sum_{\{i,j\}}\Delta^2_{i,j}\right] - \Delta^2_n.$$

Using $\mathbb{E}[\Delta_{i,j}] = \mathbb{E}[\Delta_n] = \theta$, we have

$$\begin{aligned}\mathbb{E}[S_\pi^2] &= \mathrm{Var}[\Delta_2] - \mathrm{Var}[\Delta_n] \\ &= \theta - \theta\frac{n+1}{3(n-1)} + \theta^2[f(\rho,2) - f(\rho,n)],\end{aligned}$$

and we are done. ■

Further reading

The material in this section was taken from Chapter 3 of the excellent monograph [Dur08].

References

[Dur08] Richard Durrett. *Probability models for DNA sequence evolution*. Probability and its Applications (New York). Springer, New York, second edition, 2008.