Notes 22 : Estimating the recombination rate

MATH 833 - Fall 2012 

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 22.1 (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 $\tau_a$ and $\tau_b$ 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 22.2 (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 \( \theta \) and recombination rate \( \rho(n-1)2m \) 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

\[
\Delta_n \equiv \sum_{a=1}^{m} \Delta^a_n \equiv \frac{1}{m} \sum_{\{i,j\}} \Delta_{i,j} \equiv \sum_{a=1}^{m} \frac{1}{2} \sum_{\{i,j\}} \Delta^a_{i,j},
\]

where \( \Delta^a_{i,j} \) is the number of differences between sequences \( i \) and \( j \) at locus \( a \).

Recall that

\[
E[\Delta_n] = mE[\Delta^1_n] = m \left( \frac{\theta}{m} \right) = \theta.
\]

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

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

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

\[
S^2_\pi = \frac{1}{n} \sum_{\{i,j\}} (\Delta_{i,j} - \Delta_n)^2.
\]

We will prove the following:

**THM 22.3** In the limit \( m \to \infty \)

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

where \( g \) is a function given in [Dur08].

Recall that \( \theta_\pi \) is not a consistent estimator of \( \theta \). Hence, to estimate \( \theta^2 \) we use a corrected version \( \theta^2_\pi \). This will follow from:
THM 22.4 In the limit $m \to \infty$, 

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

where

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

can be computed explicitly (see [Dur08]).

In particular, note that

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

Hence, an unbiased estimator of $\theta^2$ is

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

Putting all this together, an estimate of $\rho$ is given by the solution of

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

2 Proofs

We prove the two previous theorems. We begin with the second one.

Proof: (Theorem 22.4) Expanding the variance of $\Delta_n$, the first term gives the term not depending on $\rho$

$$\text{Var}[\Delta_n] = \sum_{a=1}^m \text{Var} \left[ \frac{1}{\binom{n}{2}} \sum_{\{i,j\}} \Delta^a_{i,j} \right]$$

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

and

$$\sum_{a=1}^m \text{Var} \left[ \frac{1}{\binom{n}{2}} \sum_{\{i,j\}} \Delta^a_{i,j} \right] \to \theta \frac{n + 1}{3(n - 1)},$$
as \( m \to \infty \), where we used (1). Rewriting the second term as

\[
\sum_{a \neq b} \text{Cov} \left[ \frac{1}{\binom{n}{2}} \sum_{\{i,j\}} \Delta_{i,j}^a, \frac{1}{\binom{n}{2}} \sum_{\{k,\ell\}} \Delta_{k,\ell}^b \right] = \frac{1}{\binom{n}{2}^2} \sum_{a \neq b} \sum_{\{i,j\}} \sum_{\{k,\ell\}} \text{Cov} \left[ \Delta_{i,j}^a, \Delta_{k,\ell}^b \right],
\]

we need to compute \( \text{Cov} \left[ \Delta_{i,j}^a, \Delta_{k,\ell}^b \right] \). By conditioning on the tree lengths \( \tau_{i,j}^a \) of locus \( a \) between \( i \) and \( j \) and \( \tau_{k,\ell}^b \), we get

\[
\text{Cov} \left[ \Delta_{i,j}^a, \Delta_{k,\ell}^b \right] = \left( \frac{\theta}{2m} \right)^2 \text{Cov} \left[ \tau_{i,j}^a, \tau_{k,\ell}^b \right].
\]

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,

\[
\sum_{\{i,j\}} \sum_{\{k,\ell\}} \text{Cov} \left[ \Delta_{i,j}^a, \Delta_{k,\ell}^b \right] = \left( \frac{\theta}{2m} \right)^2 \frac{4 \binom{n}{2}}{z^2 + 13z + 18} \left[ (z + 18) \cdot 1 + 6 \cdot 2(n - 2) + 4 \cdot \binom{n-2}{2} \right]
\]

\[
= \left( \frac{\theta}{m} \right)^2 \frac{n \binom{n}{2}}{z^2 + 13z + 18} \cdot \frac{z + (2n^2 + 2n + 6)}{z^2 + 13z + 18}.
\]

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

\[
\frac{1}{\binom{n}{2}^2} \sum_{a \neq b} \sum_{\{i,j\}} \sum_{\{k,\ell\}} \text{Cov} \left[ \Delta_{i,j}^a, \Delta_{k,\ell}^b \right]
\]

\[
= \theta^2 \frac{1}{\binom{n}{2}} \sum_{h=1}^{m} \frac{2(m - k)}{m} \frac{\rho h}{m-1} + (2n^2 + 2n + 6) \frac{\rho h}{m-1} \frac{1}{m} \left( \frac{\rho h}{m-1} \right)^2 + 13 \frac{\rho h}{m-1} + 18.
\]

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 22.3) This calculation is rather straightforward (up to a “miracle”; see [Dur08]). Rewrite

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

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

\[
\mathbb{E}[S^2_\pi] = \text{Var}[\Delta_2] - \text{Var}[\Delta_n]
\]

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

and we are done.

Further reading

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

References