

# Lecture 22 : Recombination on 2 loci

MATH285K - Spring 2010

Lecturer: Sebastien Roch

References: [Dur08, Chapter 3.1].

## 1 Recombination

The coalescence processes at two loci are correlated through a process called recombination which—looking at time going backwards—produces branchings. Formally the state of the process is described by a vector  $x = (i, j, k)$  (where  $i$  (respectively  $j, k$ ) is the number of lineages that “sit on the coalescent tree of locus  $a$  (respectively locus  $b$  and both loci)”) and the rates (going backwards) to the various states are as follows

$$(i, j, k) \rightarrow \begin{cases} (i + 1, j + 1, k - 1) & \text{at rate } r_1 = \rho k / 2 \\ (i - 1, j - 1, k + 1) & \text{at rate } r_2 = ij \\ (i - 1, j, k) & \text{at rate } r_3 = ik + i(i - 1) / 2 \\ (i, j - 1, k) & \text{at rate } r_4 = jk + j(j - 1) / 2 \\ (i, j, k - 1) & \text{at rate } r_5 = k(k - 1) / 2. \end{cases} \quad (1)$$

See [Dur08] for an illustration of the process. Letting  $n_a = i + k$ ,  $n_b = j + k$ , and  $\ell = i + j + k$ , the total rate under  $x$  is

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

## 2 A Recursion for the Covariance

To quantify the correlation between loci  $a$  and  $b$ , we consider the covariance between the total tree lengths  $\tau_a$  and  $\tau_b$ .

**THM 22.1 (Tree-Length Covariance: Recursion)** *Let  $x = (i, j, k)$  be the initial state. 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*

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

**Proof:** By the conditional covariance formula,

$$\text{Cov}[\tau_a, \tau_b] = \mathbb{E}[\text{Cov}[\tau_a, \tau_b | X]] + \text{Cov}[\mathbb{E}[\tau_a | X], \mathbb{E}[\tau_b | X]],$$

where the initial state  $x$  is implied. Let  $J$  be the time of the first jump, then

$$\tau_a = n_a J + \tau'_a,$$

where  $\tau'_a$  is the total length of the tree after  $J$ , and similarly for  $b$ . Since  $J$  is independent of  $\tau'_a, \tau'_b$  and  $X$

$$\mathbb{E}[\text{Cov}[\tau_a, \tau_b | X]] = n_a n_b \text{Var}[J] + \mathbb{E}[\text{Cov}[\tau'_a, \tau'_b | X]] = \frac{n_a n_b}{\beta_x^2} + \mathbb{E}_x[F(X)].$$

Let  $N_a$  and  $N_b$  be the number of  $a$  and  $b$  lineages after the first jump. We need to compute  $\mathbb{E}[\tau_a | X] - \mathbb{E}[\tau_a]$ . Recall that

$$\mathbb{E}[\tau_a] = h(n_a) = 2 \sum_{j=1}^{m-1} \frac{1}{j},$$

so that

$$\mathbb{E}[\tau_a | X] - \mathbb{E}[\tau_a] = \left( \frac{n_a}{\beta_x} + h(N_a) \right) - h(n_a). \quad (2)$$

Note that, considering all transitions in (1),  $n_a$  cannot increase and it decreases at rate  $r_3 + r_5$ . Hence, by a similar reasoning for  $b$ , we get

$$\begin{aligned} \mathbb{E}[\text{Cov}[\tau_a, \tau_b | X]] &= \mathbb{E} \left[ \left( \frac{n_a}{\beta_x} + h(N_a) - h(n_a) \right) \left( \frac{n_b}{\beta_x} + h(N_b) - h(n_b) \right) \right] \\ &= \frac{n_a n_b}{\beta_x^2} + \frac{n_a}{\beta_x} \frac{r_4 + r_5}{\beta_x} \left( -\frac{2}{n_b - 1} \right) + \frac{n_b}{\beta_x} \frac{r_3 + r_5}{\beta_x} \left( -\frac{2}{n_a - 1} \right) \\ &\quad + \frac{r_5}{\beta_x} \left( \frac{2}{n_a - 1} \frac{2}{n_b - 1} \right). \end{aligned}$$

Further, taking expectations in (2),

$$0 = \frac{n_a}{\beta_x} + \frac{r_3 + r_5}{\beta_x} \left( -\frac{2}{n_a - 1} \right),$$

and similarly for  $b$ , so that

$$\mathbb{E}[\text{Cov}[\tau_a, \tau_b | X]] = -\frac{n_a n_b}{\beta_x^2} + \frac{4r_5}{\beta_x (n_a - 1)(n_b - 1)}.$$

This proves the claim. ■

### 3 Solving the Recursion

The recursion in Theorem 22.1 results in linear systems that can be solved inductively in  $n_a$  and  $n_b$ . We discuss the case of 2 samples which will be useful in the next lecture.

**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.)

**Proof:** Note that  $F(i, j, k) = 0$  for  $x = (0, 0, 1)$ ,  $(1, 0, 1)$ ,  $(1, 1, 0)$ ,  $(0, 1, 1)$ ,  $(2, 1, 0)$ , and  $(1, 2, 0)$ . Hence, we get the following system of equations,

$$\begin{aligned} F(0, 0, 2) &= \frac{\rho}{\rho + 1} F(1, 1, 1) + \frac{4}{\rho + 1} \\ F(1, 1, 1) &= \frac{1}{(\rho/2) + 3} F(0, 0, 2) + \frac{\rho/2}{(\rho/2) + 3} F(2, 2, 0) \\ F(2, 2, 0) &= \frac{2}{3} F(1, 1, 1). \end{aligned}$$

This system is straightforward to solve. See [Dur08]. ■

### 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.