A Point Process Approach to Model Coalescent with Recombination

Marjan Sadeghi and Peter Beerli

Department of Scientific Computing, Florida State University

email: ms16ac@my.fsu.edu

Abstract

Wiuf and Posada (2003) introduced a coalescent model considering recombination hotspots. Their model is an extension of Hudson's coalescent model (1983), the coalescent with uniform recombination rate. They considered heterogeneity in recombination rate along the chromosome. They chose the center of recombination hotspots according to some point process and assumed that recombination happens based on a descending rate from the centers. They used a two-step procedure: using a point process to find the center of the recombination hotspot and then a chosen distribution for the recombination events happening around this hotspot. Here we propose a new model using Hawkes processes which improves on this; our model locates the recombination events and the hotspots in one step.

Hudson's Coalescent Model

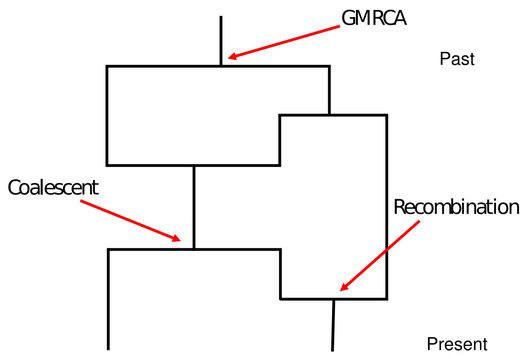

Figure 1: A simple ancestral recombination graph.

- The algorithm starts with $k = n$ sequences and looks **back in time**.
- **Time to the next event** $\sim E\big(\frac{k(k-1)}{2} + \frac{k\rho}{2}\big)$ with $\rho = 4N_e r$.
- A **coalescence** with probability $\frac{k-1}{k-1+\rho}$
- A **recombination** with probability $1 - \frac{k-1}{k-1+\rho}$
- If a **recombination** $\rightarrow$ it will be placed **uniformly** along the sequence.
- The algorithm will be repeated until the **grand-most recent common ancestor (GMRCA)** is reached.

Wiuf and Posada's Setup

They used Hudson's algorithm [2] but assumed that recombination events will be placed on the DNA non-uniformly. In what follows, we only focus on the way they evaluated recombination rate and located the hotspots. Here are the key steps they used to do this:

1. The whole chromosome is considered as a line and the region they are interested in as the interval (0,1) on this line (Figure 2).
2. The hotspots centers, x_j, $j = \pm 1, 2, ...$ are chosen according to a point process (Figure 2).
3. Recombination happens around the hotspot, x_j, with rate c_j per generation.
4. Then the breakpoint is chosen according to a distribution, g_j, around x_j (Figure 3).

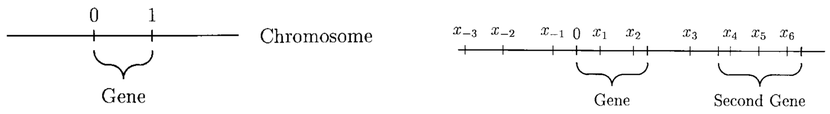

Figure 2: (Left) Chromosome and gene. (Right) The points x_j, $j = \pm 1, 2, ...$ are the centers of the hotspots.[3]

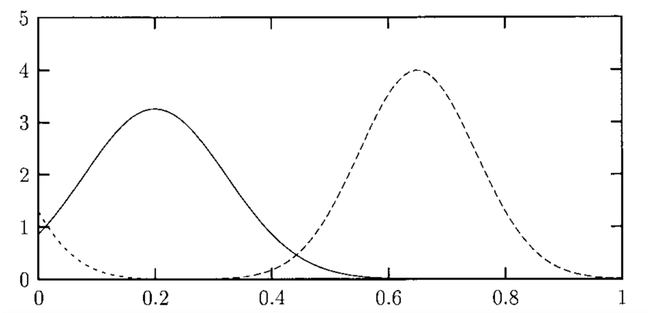

Figure 3: This shows an example of a gene that has two hotspots, one at $x_1 = 0.2$ and the other one at $x_2 = 0.65$. The breakpoints are chosen from a normal distribution, $N(0, \sigma_j^2)$ with $\sigma_j^2 = 0.015$ and 0.01, respectively.[3]

$g_j(x)$ is proportional to the probability by which recombination happens at distance x from the hotspot x_j. Therefore, the sum of the curves in Figure 3 represents the overall rate of recombination in a given point (Figure 4).

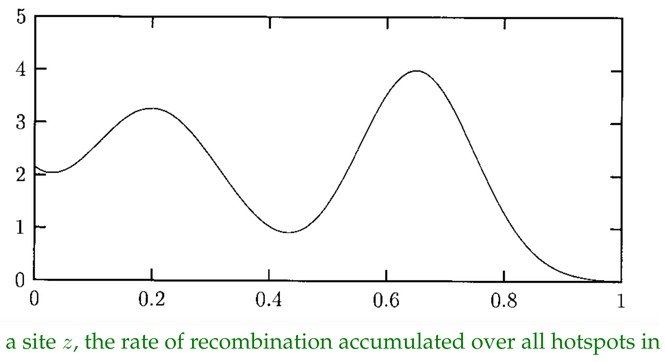

Figure 4: For a site z, the rate of recombination accumulated over all hotspots in figure 3. [3]

New Method

Our method has been setup using a category of point processes called Hawkes processes. The Hawkes processes are a class of self-exciting point processes which can be used to model the events showing the clustering effect. The intensity of this process is as follows:

$$\lambda(t) = \mu + \sum_{T_i<t} g(t - T_i)$$

in which μ is the base intensity, $g(u)$ is the excitation kernel, T_i are the occurence positions and in our case, t stands for nucleotide position [1]. The exponential kernel

$$g(u) = \alpha e^{-\beta u}$$

is one of the common type of the excitation kernels for Hawkes processes in which α is the *size of self excited jumps* and β is the *exponential decay rate* and $\alpha, \beta > 0$. Figure 5 shows a simulation example of a Hawkes process with intensity function $\lambda(t) = 0.3 + \sum_{T_i<t} 0.5e^{-1.5(t-T_i)}$.

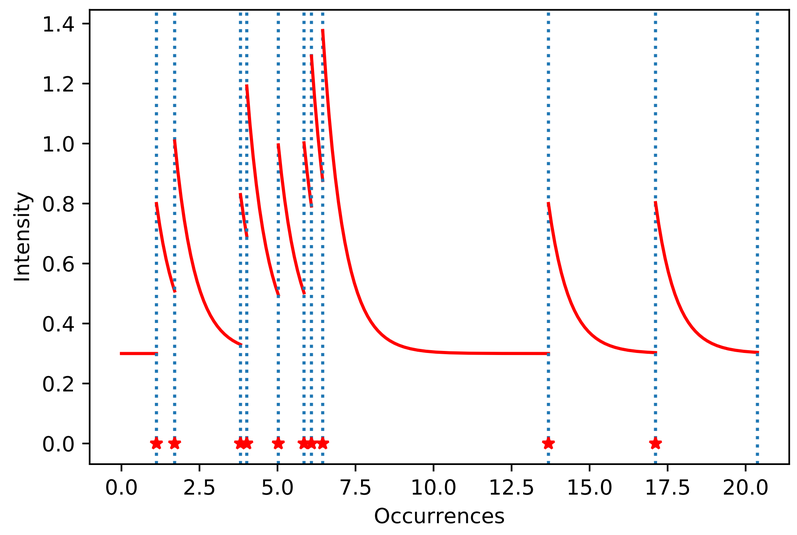

Figure 5: Simulation of a Hawkes process with exponential kernel and the intensity function $\lambda(t) = 0.3 + \sum_{T_i<t} 0.5e^{-1.5(t-T_i)}$ on interval [0, 20]. The red dots shows the simulated data which are the occurrences of the Hawkes process.

Each time there is an event the intensity will jump by α. The parameter β controls how fast the intensity function decays in each interval. The bigger the β, the faster the decay.
Considering the exponential kernel and its properties, we developed a new Hawkes process to model the recombination events along the chromosome. The intensity function of our model is defined as follows

$$\lambda(t) = \mu + \sum_{T_i<t} \alpha_i e^{-\beta_i(t-T_i)}$$

We used different parameters of α_i and β_i to increase the accuracy of a simple Hawkes. The Figure 6 shows the intensity of our model along the DNA.

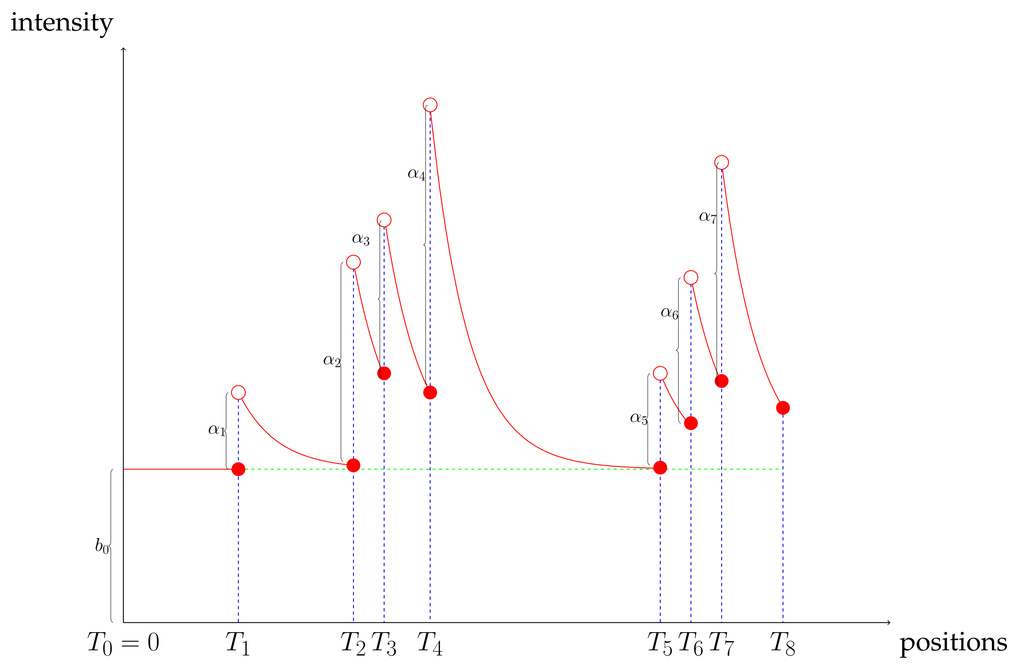

Figure 6: The Hawkes process with the intensity function $\lambda(t) = b_0 + \sum_{T_i<t} \alpha_i e^{-\beta_i(t-T_i)}$. The parameters β_i control the decaying rate.

The intensity function in this model is considered as the recombination rate. This process is able to cover the clustering effect of the recombination events. Therefore, it will be able to locate the hotspots as well. Hence, our model replaces all the four steps in Wiuf and Posada's algorithm for evaluating the recombination rate and locating the hotspots by one step.

References

[1] Hawkes, A.G., 1971. Spectra of some self-exciting and mutually exciting point processes. Biometrika, 58(1), pp.83-90.

[2] Hudson, R.R., 1983. Properties of a neutral allele model with intragenic recombination. Theoretical population biology, 23(2), pp.183-201.

[3] Wiuf, C. and Posada, D., 2003. A coalescent model of recombination hotspots. Genetics, 164(1), pp.407-417.