Modeling Recombination Events Using Hawkes Processes Along Sequences

Marjan Sadeghi and Peter Beerli

Florida State University, Department of Scientific Computing

email: ms16ac@my.fsu.edu

Abstract

Haldane introduced the idea of using Poisson point processes to model recombination events in 1919. Recently, this process was incorporated in the genealogical history of a sample of sequences to handle recombination and coalescence, for example, Wiuf and Hein (1999) [6]. The Poisson process assumes that recombination events occur independently, and hence ignores phenomena such as interference and hotspots. We propose to replace Poisson processes with Hawkes processes to model recombination events. Poisson process has a deterministic intensity function, while a Hawkes process has stochastic intensity function making the occurrences of events depend on the entire history of the process. Therefore, the proposed algorithm will cover the dependency between the recombination events, such as interference and hotspots. We hope our algorithm improves the accuracy of recombination estimation.

Introduction

In 1999, Wiuf and Hein [6] developed a spatial algorithm to model the history of a sample of sequences considering the recombination events to be a Poisson point process along the DNA sequences.

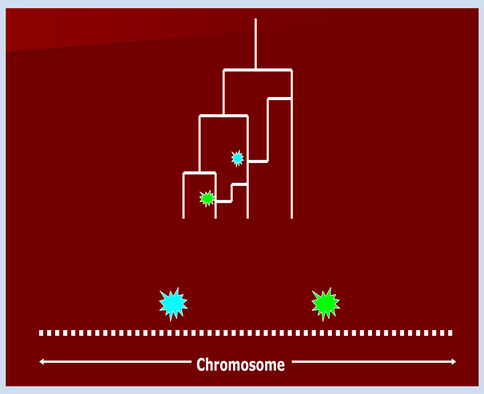

Figure 1. Wiuf and Hein algorithm. [4]

Their algorithm was slow, therefore, many other algorithms have been developed by some modifications of their original idea to improve the speed among which the Sequentially Markovian coalescent process (SMC) methods ([5], [4]) are popular.

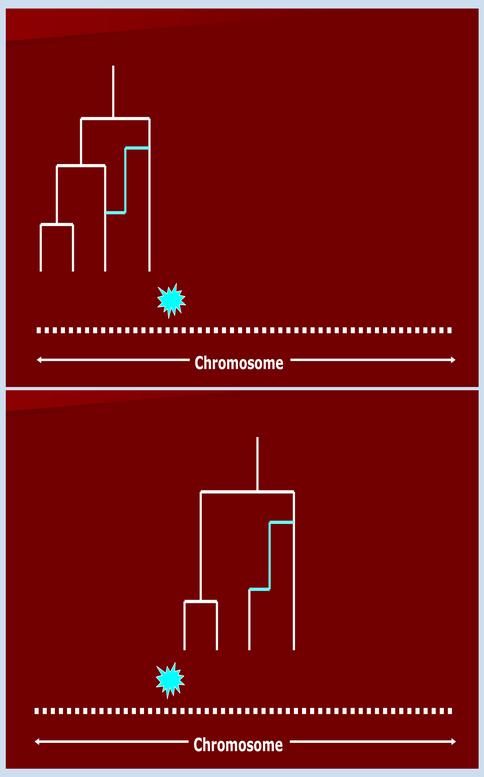

Figure 2. SMC algorithm. [4]

SMC methods are faster, however, they looses the accuracy.
We are going to present a new method using a different point process to provide more accurate results.

General Approach

The general approach of the existing algorithms is considering the total number of recombination events in a segment length of DNA, conditional on the total branch length b of the genealogy, to have a Poisson distribution. Therefore, the chromosome length X to the first event will follow an exponential distribution as:

$$\mathbf{P(no\ recombination \mid b) = exp(-\frac{\rho}{2}b)}$$

Events in a Poisson process occur independently. Therefore, all of these methods ignore the dependency between the recombination events such as **interference** and **hotspots**. Hence, in many cases, the algorithm will not be able to model the events accurately. The following figures prove this fact.

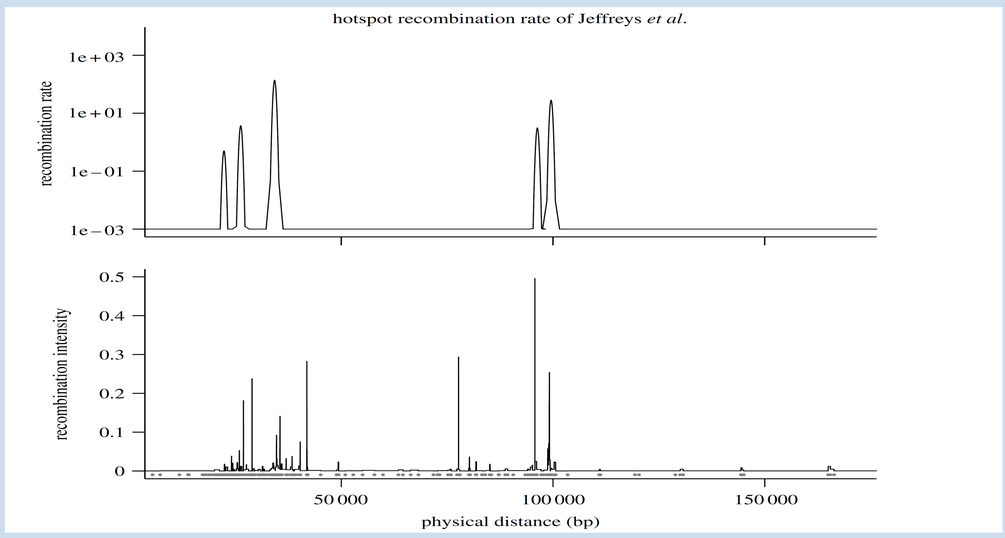

Figure 1. Top figure shows recombination rates inferred from sperm-typing studies and the bottom shows the statistically inferred recombination intensities for the HLA. [2]

The figure shows an inconsistency between the experimental recombination rate and the one inferred from the simulation studies. The experimental rate shows a **self-exciting clustering** effect which can not be covered by the constant rate of a Poisson process.

New Method

The **Hawkes process** is one of the most famous self-exciting point processes; its intensity is defined as

$$\boldsymbol{\lambda_t} \quad = \quad \boldsymbol{\mu} + \int_{\mathbf{s<t}} \mathbf{g(t-s)\,dN_s} \quad = \quad \boldsymbol{\mu} + \sum_{\mathbf{T_i<t}} \mathbf{g(t-T_i)},$$

where t stands for nucleotide position, $\mu > 0$ is the *base intensity*, $g(u) : \mathbb{R}^+ \to \mathbb{R}^+$ is the *excitation kernel* of the process, N is the associated *counting process*, and T_i are the occurrence positions [3]. This process is used to model the events with clustering features.
We are trying to use an adapted version of the Hawkes process [1] to model recombination event:

$$\boldsymbol{\lambda_t} \quad = \quad (\boldsymbol{\mu} + \int_{\mathbf{(-\infty,t)}} \mathbf{g(t-s)\,dN_s})^{+}$$

where $g(u) : \mathbb{R}^+ \to \mathbb{R}$. We define $g(u)$ as

$$\mathbf{g(u) = \alpha_1 \exp{(-\beta_1 u)} - \alpha_2 \exp{(-\beta_2 u)}}$$

in which $\alpha_1, \beta_1, \alpha_2$ and β_2 are the parameters of the model need to be calibrated. The adapted version is able to cover the **self-excitation** and **self-inhibition** features of the recombination events at the same time. Therefore, our model will be able to cover **interference** and **hotspots** phenomena. Hence, we expect our algorithm produce more accurate results than the existing methods.

References

[1] Costa, M., Graham, C., Marsalle, L. and Tran, V.C., 2018. *Renewal in Hawkes processes with self-excitation and inhibition*. arXiv preprint arXiv:1801.04645.

[2] De Iorio, M., de Silva, E. and Stumpf, M.P., 2005. *Recombination hotspots as a point process*. Philosophical Transactions of the Royal Society B: Biological Sciences, 360(1460), pp.1597-1603.

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

[4] Marjoram, P. and Wall, J.D., 2006. *Fast "coalescent" simulation*. BMC genetics, 7(1), p.16.

[5] McVean, G.A. and Cardin, N.J., 2005. *Approximating the coalescent with recombination*. Philosophical Transactions of the Royal Society B: Biological Sciences, 360(1459), pp.1387-1393.

[6] Wiuf, C. and Hein, J., 1999. *Recombination as a point process along sequences*. Theoretical population biology, 55(3), pp.248-259.