

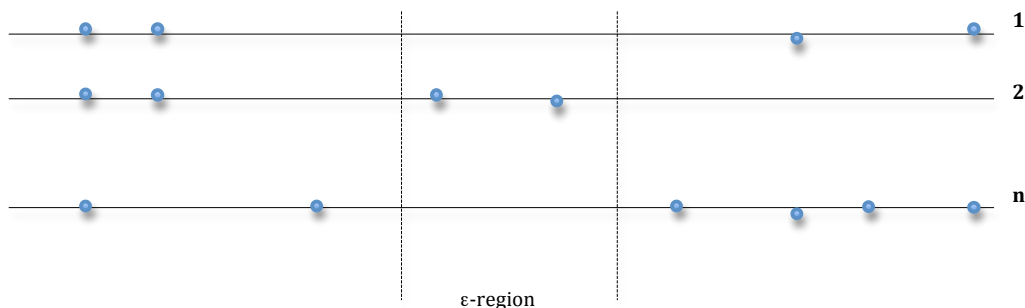
A Gibbs Sampler of the Ancestral Recombination Graph

15.7.08

Motivation and Background. DNA sequences from a population has an unobservable genealogical history. In analyzing such data, a crucial stepping stone is to be able to integrate over evolutionary histories according their probability according to a model and given the data. Doing this has been the focus of research for more than 2 decades. Doing this has been the focus of research for more than 2 decades. The basic probability model for genealogical histories with recombination was given Watterson (1975) and Kingman (1982). Until 1994 (Griffiths and Tavaré), this was solely used as a tool for simulating genealogical histories without knowing the content (mutational configuration) of the sequences. Since late 90s there has been a string of attempts to apply stochastic integration methods (Importance Sampling, MCMC,...) to do this. In the absence of recombination it is a hard, but doable problem. Due to the enormous increase in DNA sequences from populations and the importance of this problem in genetic mapping, the problem remains as important as ever.

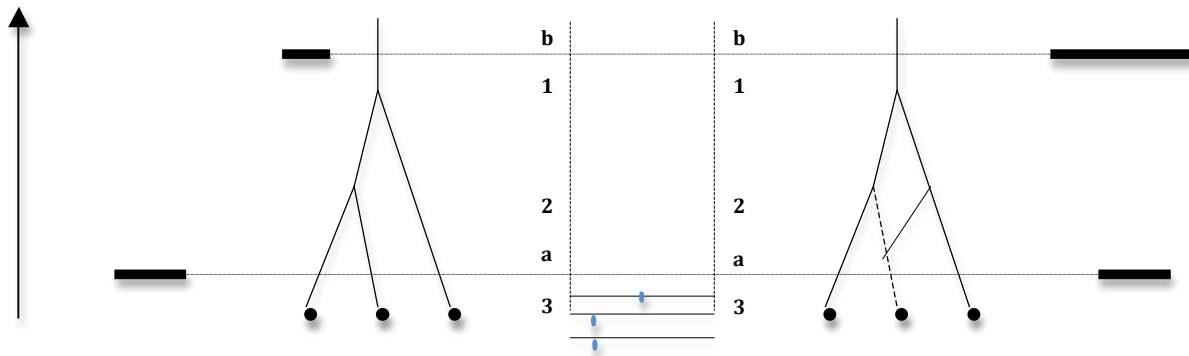
Recursions (Ethier and Griffiths, 1987) can be written to calculate the likelihood of a data set in terms of possible ancestral configurations (Song, Lyngsø and Hein, 2006). The number of ancestral configurations grows as a function of number of sampled sequences and segregating sites but also depends on the exact configuration of segregating sites. Griffiths and Marjoram (1996) and Fearnhead and Donnelly (2001) devised importance samplers to intergrate of histories. Lyngsø, Song and Hein (2008) used a branch and bound algorithm to analyze a model with recombination. Presently there exists no satisfactory method to do this for decent data sizes.

Project Description. A standard Gibbs sampler (GS) will take a vector (k dimensional) valued stochastic variable X and sample the value of X with all indices but one (say i) fixed. Then chose a new index etc. There are many, many variants of this procedure. This can be applied to the ancestral recombination graph (ARG), which is index by a continuous variable (position) that lies in an interval specified by the length of the sequences measured in normalized recombination rate r . One way to apply the Gibbs sampler technique would be to chose a segment on the interval and then remove all linkages between histories to the left and to right of this interval and then sample the histories of the of the interval. See first illustration. The structure of this Gibbs sampler has some resemblance to the one introduced by Jensen and Pedersen (2000) in a model of sequence evolution that incorporated neighbour dependence.



One Gibbs sampling step will take a window of with width ϵ and “forget” its history. The history of the sequences in this window will then be resampled conditional on the histories of the sequences in the regions flanking the window. The window will be moved back and forth along the sequences until mixing has been achieved.

However, there are some novelties in this GS relative to the standard one. The separation of the history into what happens in the ϵ -region and outside is not that clear cut (see second figure). What happens in the ϵ -region, affects the histories outside. Basically a recombination event in the ϵ -region will unlink sequences on each side of the ϵ -region. Since sampling histories of a segment is very difficult for longer segments, this should probably be done for segments, where the expected number of recombinations are 0 or 1. This can be solved using reasoning in the EGT algorithms and in the paper by Lyngsø et al. (2008), but sequences can enter further back in time, in contrast to the present program, where all sequences appear at the present. These sequences appearing further back in time is due to coalescent events between ancestral segments on each side of the ϵ -region.



In the present we have observed 3 sequences, now illustrated as black balls. At the left side of the ϵ -region there will be local tree relating the sequences at that point and correspondingly at the right point. We will rerun history for the ϵ -region and start at the present and go back in time. However, the histories outside the ϵ -region places constraints on what can happen in the ϵ -region. At the present we have the observed sequences. As we trace them back in time, they will have to follow the histories to the sequences they are linked to outside the ϵ -region, except if a recombination happens to them in the ϵ -region. Such sequences are indicated by **a** and **b** at the border regions of the ϵ -region. Then the histories of the sequences to the left and right of the recombination point becomes unlinked. As we go back in time in the ϵ -region, sequences linking ancestral material outside the ϵ -region will sudden appear and their history will have to be re-played as well.

Plan

- Read Hein, Schierup and Wiuf (2005) chapt 1-3 + 5
- Implement Hudson (1983) algorithm in a program, HUDSON.
- Use the program from Lyngsø et al. (2008) to calculate the probability of small data sets experiencing recombination.
- Modify this program, so it can also take non-extant sequences call it ϵ -REC.
- Combine ϵ -REC and HUDSON to do the Gibbs sampler approach.

Comments. i. This project would need reading a few papers, but mainly programming skills. ii. The full conditional distribution in the ϵ -region is known, but is only used to sample a single history. Maybe this could be done more efficiently.

References.

- Ethier and Griffiths (1987) "The infinitely many sites model as a measure valued diffusion" *Ann. Prob.* 15.2:515-545.
 Griffiths, R.C. (1989). "Genealogical-tree probabilities in the infinitely-many-sites model". *J. Math. Biol.* 27, 667-680.
 Griffiths, R.C. and Tavaré, S. (1995). "Unrooted genealogical tree probabilities in the infinitely-many-sites model". *Mathematical Biosciences* 127, 77-98.
 Griffiths, R.C. (2001). "Ancestral inference from gene trees" In: Donnelly, P. and Foley, R. (Eds.), *Genes, Fossils, and Behaviour: an Integrated Approach to Human Evolution*, IOS Press, Amsterdam, pp.137-172.
 Hein, Schierup and Wiuf (2005) "Gene Genealogies, Variation and Evolution" OUP
 Jensen, J.L and Pedersen, AMK (2000) Probabilistic Models of DNA Sequence Evolution with Context Dependent rates of Substitution" *Adv. Appl. Probab.* 32:499-517
 Liu, J. (2001) Monte Carlo Strategies in Scientific Computing Springer
 Lyngso, Song and Hein (2008) "Accurate calculations of likelihoods in the coalescent with recombination using parsimony" *Recomb 2008* Singapore
 Fearnhead, P. and Donnelly P. (2001) "Estimating recombination rates from population genetic data", *Genetics* 159.3:1299-1318.
 Griffiths, R.C. and Tavaré, S. (1994). Simulating probability distributions in the coalescent. *Theor. Popul. Biol.* 46, 131-159.
 Griffiths, R.C. and Marjoram, P. (1996). "Ancestral inference from samples of DNA sequences with recombination" *Journal of Computational Biology* 3, 479-502.
 Hudson, R.R. (1983) "Properties of the neutral model with intragenic recombination" *Theor.Pop.Biol.* 23.2:213-201.
 Li, N., and Stephens, M. (2003). Modelling Linkage Disequilibrium, and identifying recombination hotspots using SNP data *Genetics*, 165:2213-2233.Liu, .()
 Song, Y, R.Lyngsø & J.Hein (2006) "Counting Ancestral States in Population Genetics" *Bioinformatics and Computational Biology* vol.3.3:239-252
 Stephens, M. and Donnelly, P. (2000). Inference in Molecular Population Genetics. *Journal of the Royal Statistical Society, Series B*, 62, 605–655