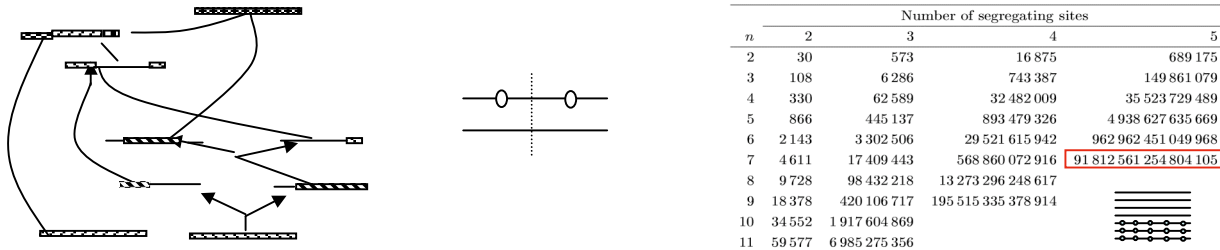


Workbench for Ancestral Recombination Graph summation

1.3.08

The *Ancestral Recombination Graph* (ARG) is the graph that describes the relationship of a set of homologous sequences subject to recombination. It was first introduced by Griffiths (1981) and Hudson (1983). See illustration below for the two basic events in an ARG – the coalescent and the recombination event. An evolutionary history can be translated into an ARG by starting in the present and going backwards in time until all positions of the sequences have found one single ancestor. Going back in time, sequences encounter coalescences and recombinations. Coalescent events will merge sequences that are identical, reducing the sample size by one. Recombinations will redistribute a single sequence to two sequences, where one sequence will carry the material to the left of the recombination point and the other the material to the right of that point. In most analysis the ARG ignores that sequences are in individuals and thus describes a population of sequences, not of individuals with sequences.

An ARG is a sequence of recombinations and coalescent events starting in the present with a sample of sequences and going backward in time until all points at the same position in the sample have found a common ancestor.



Left. Modified from Hudson (1991). Starting in the present with 2 sequences at the bottom of the illustration and going back in time by going upwards in the figure. The most recent event is that the leftmost sequence was created by a recombination of two sequences. The sequence segments that are not ancestral to any of the two extant sequences are represented by thin lines. This ARG has 2 recombination events and 2 coalescent events. At the very top (most ancient) is a sequence that is ancestral to both sequences in the bottom.

Middle. Two sequences, one with 2 mutants on it (balls), the other with no mutations on it. All recombination events have been forced into the point at the middle between the two segregating sites. The number of possible ancestral states will then be finite and configurations will have probabilities and not densities.

Right. The number of possible ancestral states in an ARG as a function of sequence number and segregating sites. Clearly very very large numbers, prohibiting exhaustive investigations. The middle configuration can visit 30 possible ancestral states.

It is clear from the size of this problem that computational approaches that calculating the likelihood cannot be based on using the full recursions. It is also clear that very very large accelerations are needed as the factor $10^6 \cdot 10^9$ was not enough in the Lyngsø paper. A series of approaches can be envisioned.

i. **Upper and Lower Bounds.** Lyngsø's approach of discarding unlikely configurations clearly gives a lower bound and probably a very tight one. There is presently no useful upper bound. It is conceivable that graphs could be defined by lumping states – ARG^{upper} and ARG^{lower} – the would allow calculation of upper and lower bounds on the likelihood.

ii. **Symmetries and Precalculations.** The ARG is defined by the same events over and over and has many inherent symmetries. However, they are very hard to put to efficient use as the ARG summation involves products and summations and symmetries are often in different contexts or only almost symmetries. Precalculation can be done when the same calculation is done again and again and is a standard compsci trick and could most likely be applied, but will hardly give more than a single order of magnitude.

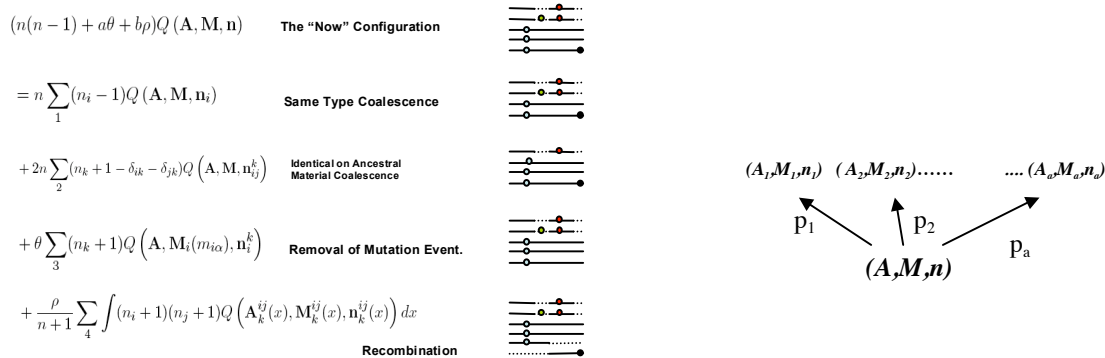
iii. **Inclusion-exclusion** arguments. Since we can calculate the probability of low cardinality configurations, it would be natural to make statements about high cardinality configurations, but picking out suitable sets of low cardinality configurations and combine to make upper and lower bounds.

iv. **Stochastic Integration – MCMC and IS.** Most effort has gone into this approach, partly due to the phenomenal rise of computationally intensive statistics. It is the aim of this project to evaluate these methods.

v. **Approximating Genealogical Structures.** Such an approach is inherent in Li and Stephens (2001) and has been of great practical value. Such approaches can also be intellectually frustrating as it can be difficult to interpret results in terms of sequence histories, which is central to coalescent theory.

vi. **Heuristics** can be very useful and in many situations the only way forward.

The most famous importance samplers have been developed by Griffiths and Marjoram (1996) and Fearnhead and Donnelly (2001). The fundamental approaches is the following: If $Q(A, M, n)$ is the probability of a configuration and $\{Q(A', M', n')\}$ the set of precursors and $q(A, M, n, A', M', n')$ the weight in the recursion below, then $Q(A, M, n)$ can be estimated by . GM96 chose ??, while FD00 chose ??,



Left. A probabilistic recursion can be formulated. The present configuration (top lines) can be decomposed into which configurations could have created it in one event.

Right. The equation can be solved by choosing random histories starting at the present configuration and going back to a certain starting state. It must be chosen at each step with which probability to choose among the possibilities. The probability of each path must be inflated so the expectation corresponds to the probability summing over all paths.

Project:

Lyngsø has written a program – COB - that can solve the equations systems (almost) exactly for up to 7 sequences with 7 segregating sites dependent on ρ and θ . “Almost” as the program will have to ignore configurations that contribute negligibly to the likelihood. Although this is far from real data, it allows investigations of the performance of how non-exact methods fare relative to this exhaustive investigation. The project should be done in collaboration with Lyngsø as there are extensive programming and data structures that can be re-used in COB.

What are natural indicators measuring how importance samplers fail/succeed?

Plan:

- i. Implement or use the recursions/program for the situation without recombination as described in “Corner-cutting approaches to the EGT recursions”.
- ii. Implement and test the GT94 sampler for the EGT recursions.
- iii. Implement GM96.

References:

Fearnhead, P. and Donnelly P. (2001) “Estimating recombination rates from population genetic data”. *Genetics* 159.3.1299-1318.
 Griffiths, R.C. (1981). Neutral two-locus multiple allele models with recombination. *Theor. Popul. Biol.* 19, 169-186
 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.
 Hein, J.J., Schierup, M.H. and Wiuf, C.H. (2005) *Gene Genealogies, Variation and Evolution*. Oxford University Press
 Hudson, RR (1991) “Gene Genealogies and the Coalescent Process” *Oxford Surveys in Evolutionary Biology* 7.1-49
 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

Comments. i. This project is probably one of the most difficult ones on the project page. The underlying problem is hard, the papers are hard and it will take good coordination with COB (Lyngsø) to perform well. However, it is of central importance to the field and is of great value to understand in depth for anybody interested in association mapping and population genetics, so even a valiant attempt of progress might worth it. ii. It is encouraged that a report is written continuously on progress and results. The project should appeal to anyone, who likes combinatorics, algorithm and programming. It can be addressed very directly without much literature study, but could also involve studying the original literature on genealogical structures and combinatorial counting.