

## Counting Ancestral Recombination Graph (ARG) Topologies

In tracing or describing the history of species, individuals in a bi-sexual population and sequences subject to recombination are used 3 graphs: The phylogeny, the pedigree and the ancestral recombination graph. These graphs have both a continuous aspect and a discrete aspect. The continuous aspect would describe branch lengths and dates that can be parameterized with real numbers. The discrete aspect is often called “the topology” by biologists. Counting this, has been done by Cayley and Prufer in non-biological contexts. Felsenstein (1978) counted a series of elementary cases. Griffiths (1987) counted a case that arises in population genetics. Counting pedigrees have been done in restricted cases by Thomas and Cannings (2003). A start of counting pedigrees was done in a 2<sup>nd</sup> year 6-week project, that can be found at <http://mathgen.stats.ox.ac.uk/bioinformatics/projects/>, but much more remains to be done. Counting the last combinatorial structure – the ARG – seems to be virgin territory. The number of topologies of different genealogical structures is useful both as a measure of the hardness of problems involving that structure, but can also be a useful stepping stone towards a better understanding of the problem.

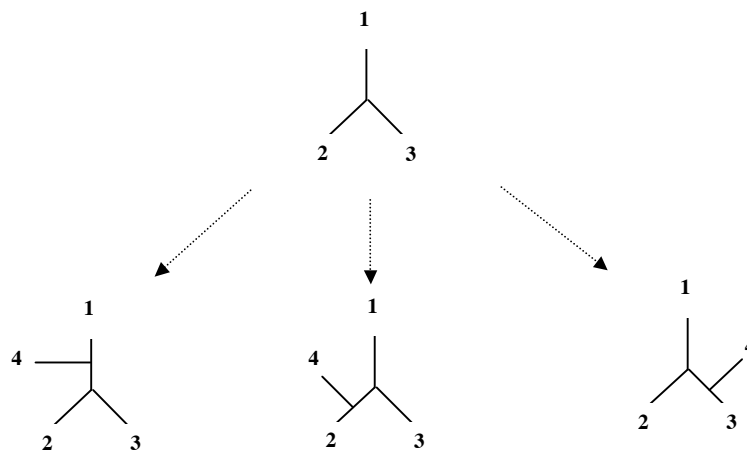
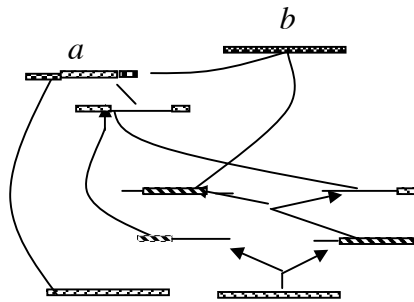


Illustration of simple recursive counting argument. There is only one unrooted tree topology with three labelled leaves and unlabelled inner node. The corresponding number with 4 leaves and only having duplications at inner nodes must be 3 as there are 3 edges at which the 4<sup>th</sup> leaf can be added. This argument allows tree counting for larger number of leaves.

The *Ancestral Recombination Graph* (ARG) is the graph that describes the relationship of a set of homologous sequences subject to recombination (Hein, Schierup and Wiuf, 2005). 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.



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.

This figure also illustrates the simple fact that each position in the sample sequences are related by a traditional phylogeny, but that this phylogeny can change as this position is moved along the sequences. Above the very ends of the sample sequences finds an ancestor at the sequence labelled *a*, while the middle segment will find a common ancestor at the sequence labelled *b*.

#### Project:

Just counting ARG topologies is very general as there are many restrictions, so a few specifications are in place and generalisations can easily be formulated.

Let the sequences in the present be labelled 1,...,n. Only duplications (ie no triplications or beyond) occur in their history. All the events (recombinations and coalescents) in the ARG will be ordered.

- Count the number of ARG topologies with k recombinations.
- Count the number of k local trees.
- How many ARG topologies will realize a given sequence of k local trees. What is the sequence of local trees maximizing/minimizing this number?

Counting can be done very basically by paper and pen for a start and slightly more advanced by dynamic programming scanning along the sequences or going backwards in time. It is unclear how far the dynamic programming approach can count as it is conceivable that simplifications in the problem can be found.

Other approaches that could be attempted would be probabilistic algorithms and finding upper and lower bounds on the numbers. Characterizing the asymptotic growth is also of interest.

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.

- Felsenstein, J (1978) The Number of Evolutionary Trees *Systematic Zoology*, Vol. 27, No. 1. 27-33.
- Griffiths, R.C. (1981). Neutral two-locus multiple allele models with recombination. *Theor. Popul. Biol.* 19, 169-186
- Griffiths, R.C. (1987). Counting genealogical trees. *J.Math.Biol.* 25, 422-432.
- Hein, J.J., Schierup, M.H. and Wiuf, C.H. (2005) *Gene Genealogies, Variation and Evolution*. Oxford University Press, 296 pages. chapter 5
- 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.
- Semple and Steel (2003) "Phylogenetics" OUP
- Song, Y, R.Lyngsø & J.Hein (2006) "Counting Ancestral States in Population Genetics" *Bioinformatics and Computational Biology* vol.3.3.239-252
- Thomas, A. and Cannings, C. (2003) Enumeration and simulation of marriage node graphs on zero loop pedigrees. *Math. Med. Biol.* 20: 261-275.
- van Lint and Wilson (1992) *A Course in Combinatorics* CUP (chapter 2. Not easy book – covers material at fast pace)