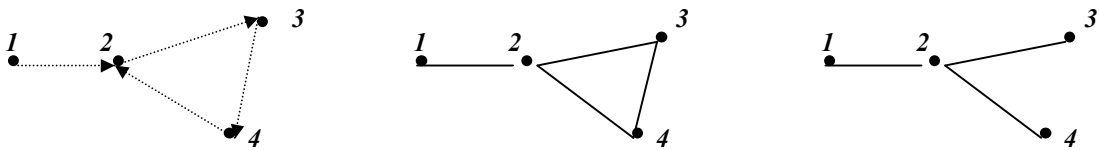


# Counting in Phylogenetics

1.7.07

Phylogenetics is the science of the relationship of objects (sequences, individuals,..) that can be viewed as having an evolutionary history. The most famous of this is undoubtedly the tree (phylogeny) relating the species on Earth – the Tree of Life (TOL).

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. A graph consists of nodes and edges. A graph can have  $k$  nodes  $\{n_1, n_2, n_3, \dots, n_k\}$  and  $l$  edges are pairs of nodes that are connected  $\{(n_{1_1}, n_{2_1}), (n_{1_2}, n_{2_2}), \dots, (n_{1_i}, n_{2_i})\}$ . Edges can be directed or undirected. If directed one can envision an arrow from the first node to the second node. If the edge is undirected the first and second node can be permuted without changing the graph.



The leftmost graph is directed and has nodes  $\{1,2,3,4\}$  and  $\{(1,2),(2,3),(3,4),(2,4)\}$ . The middle graph is the same except that is undirected. The rightmost is the same as the middle, except the edge  $(3,4)$  has been removed. The rightmost graph is then a tree since it connected (you can find a path from any node to any node) and there are no cycles (you cannot find a series of edges, that bring you back to the same node). The valency of a node is the number edge touching it. A leaf is only touched by one node.

The graphs in phylogenetics 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.

Phylogenies can be root if there is a specific node that is the most ancient in the tree. Phylogenies often has labelled leaves and unlabelled non-leaf nodes (internal nodes). If we want to count unrooted tree topologies, where internal nodes have valency 3, it is clear that there is only one with 3 labelled leaves. If we let  $T(k)$  be the number of tree topologies with  $k$  leaves and internal valency 3. Thus  $T(3)=1$ .

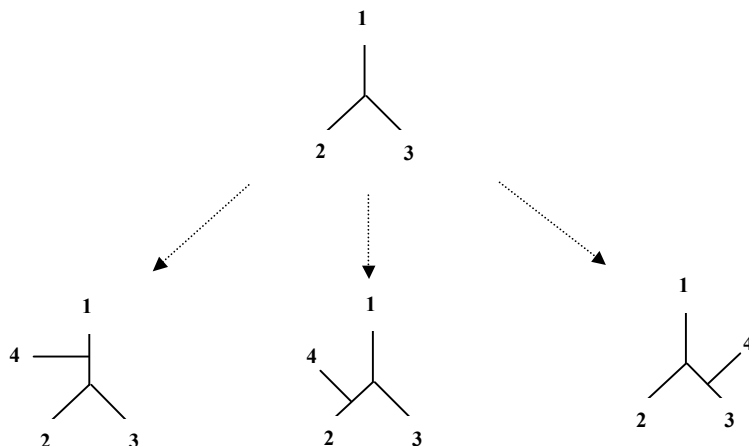


Illustration of simple recursive counting argument. There is only one unrooted tree topology with three labelled leaves and unlabelled inner node. The corresponding number with 4 leaves and only having duplications at inner nodes must be 3 as there are 3 edges at which the 4<sup>th</sup> leaf can be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

It is easy to see that an unrooted tree topology with internal nodes valency 3 and  $k$  leaves, has  $2k-3$  edges. Thus in adding a  $k$ 'th leaf to a unrooted tree topology with internal nodes valency 3 and  $k-1$  leaves, there are  $2k-5$  edges, where to add this last leaf. Ie  $T(k)=(2k-5)T(k-1)$ . This recursion allow tabulation of the number of possible topologies and in this simple case even a closed formula:  $T(k)=(2k-5)!/[(k-3)!2^{k-3}]$ .

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.

### **Project Plan:**

Week 1-2:

- Learn PYTHON
- Write simple program that will count unrooted phylogenies with valency 3.

Week 3-4. Write recursions that the counts and write programs for all of these situations.:

- any unrooted phylogenies with arbitrary degree on internal nodes
- with bounded degree on internal nodes and with labelled internal nodes.
- Tree topologies if internal nodes are ranked?

Week 5-6: Address more open ended problems:

- What is the asymptotic growth for all the above cases ie find continuous functions –  $f(t)$  - that has the same growth as the number of topologies for different cases.
- How many topologies are there if 1,2,3, recombination events are allowed.

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.

### **References**

- Aigner and Ziegler (2003) "Proofs from the Book" 3<sup>rd</sup> ed. Springer Chapter: chapter 22 "Cayley's Formula .."
- Felsenstein, J (1978) The Number of Evolutionary Trees *Systematic Zoology*, Vol. 27, No. 1. 27-33.
- 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.
- Semple and Steel (2003) "Phylogenetics" OUP
- <http://www.tolweb.org/tree/phylogeny.html>
- <http://www.cs.uwaterloo.ca/journals/JIS/cayley.html>
- van Lint and Wilson (1992) *A Course in Combinatorics* CUP (chapter 2. Not easy book – covers material at fast pace)

**Discussion times:** 16-18-20.7 with Jotun Hein always 8.30AM. 23-25-27-30.7 with Joanna Davies. 1-3-6-8-10-13.8 with Rune Lyngsø. 15-17-20-22-24.8 with Jotun Hein. These times are only there to make sure that we discuss with you – you would in general be welcome to come by any time.