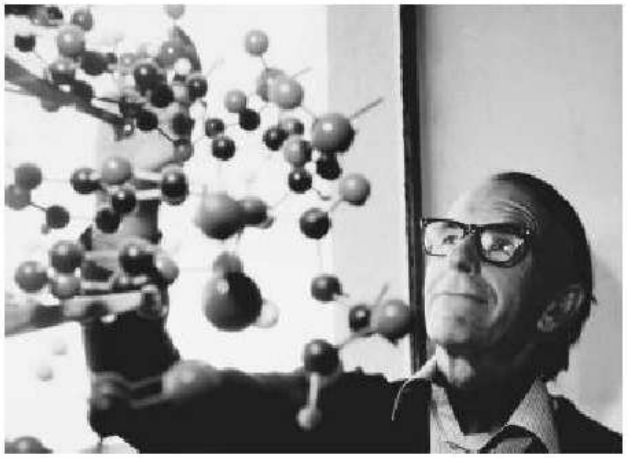
Statistical population genetics

Lecture 8: Infinite sites model

Xavier Didelot
Dept of Statistics, Univ of Oxford
didelot@stats.ox.ac.uk

Infinite sites model

- Watson and Crick (1953) discovered the **structure of DNA**.
- Genes are made of a **sequence of nucleotides** which can take four values: A, C, G and T.
- Sanger (1975) found a way to **sequence DNA**.
- This gives **more information** than electrophoresis.
- The infinite alleles model is **still usable**, but does not take advantage of the availability of gene sequences.



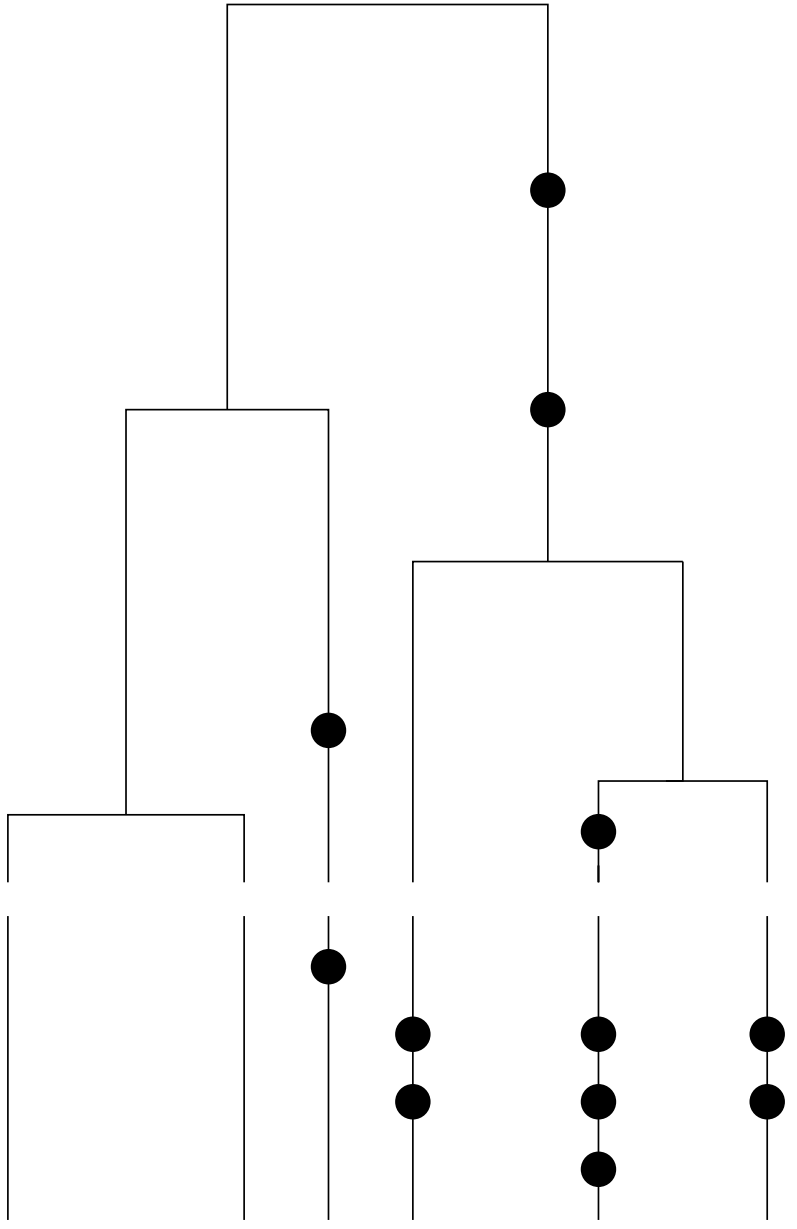
Infinite sites model

- These facts prompted Kimura (1969, 1971) and Watterson (1975) to introduce the following mutational model:

  Definition (Infinite sites model)**.**
  Each gene is made of a sequence of sites, and each time a mutation occurs, it affects a site that was previously unaffected.

- In the infinite sites model, **all mutations have an impact** on the data, unlike the infinite alleles model
- The infinite sites model is still only an **approximation** to reality, because the same site can be affected more than once in reality

Infinite sites model



Infinite sites model

- The number of **mutant sites** (also called **polymorphic sites**) in the infinite sites model is equal to the number S_n of mutations occurring on the coalescent tree
- The distribution of S_n has been previously calculated
- Data under the infinite sites model can be represented as an **incidence matrix** where the columns represent the polymorphic sites and the rows the sequences.
- A 0 indicates identity with the sequence of the MRCA and a 1 a difference.
- On the previous slide, the incidence matrix is:

$$M = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

- The order of columns and rows in the incidence matrix is arbitrary.

Gene trees

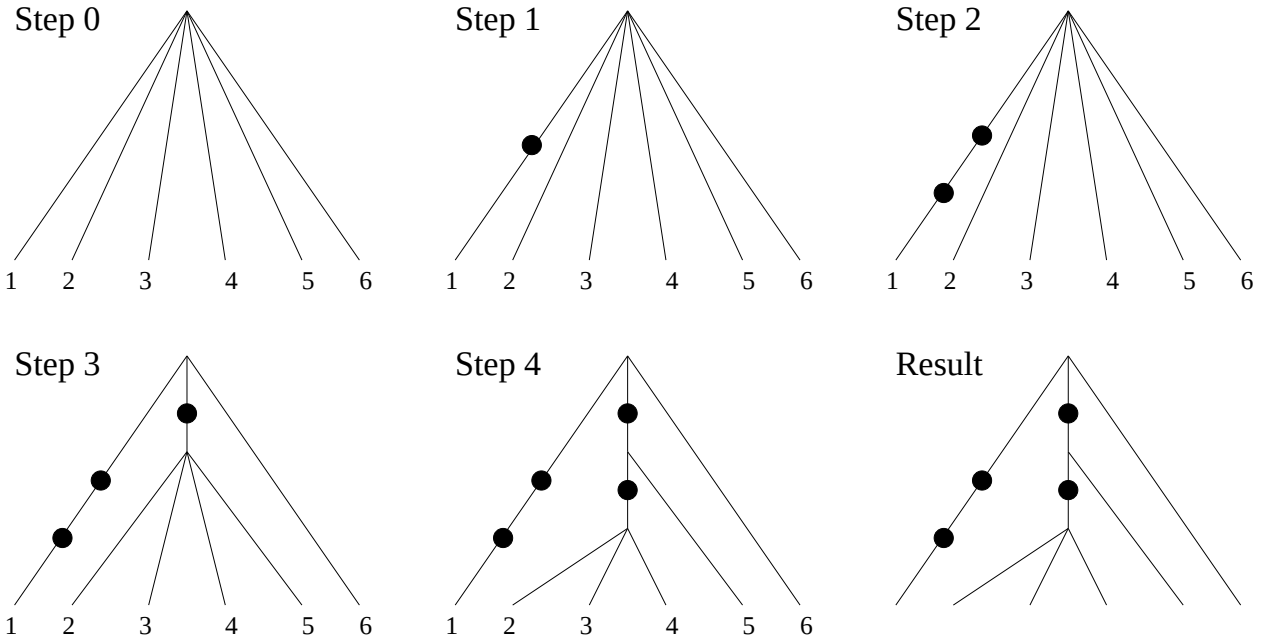
Definition (Gene tree)**.**

A gene tree is constructed from an incidence matrix as follows:

- *Start with a leaf per sequence, a tree root, and a branch linking each leaf directly to the root. Give each branch the label 0;*
- *Consider each mutation in turn and the set S of sequences that have the mutation.*
- *If there is a branch in the tree subtending exactly the leaves corresponding to the sequences in S, increase its label by one.*
- *Otherwise, find the node m which subtends the smallest superset of S. Create a new node n with father m and label 1 on the branch between n and m. Regraft under n the children of m which subtend elements of S (and the branches above those children keep their labels).*

Example

Construction of the gene tree corresponding to the matrix $\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$:



Example

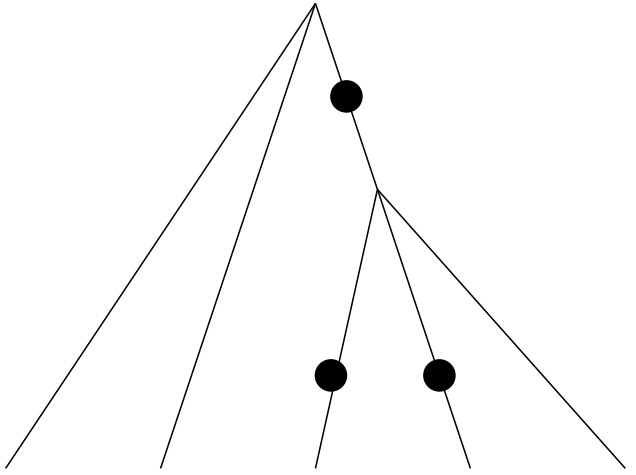
- The branch labels are indicated by the number of circles on each edge
- The labels for the leaves are shown to make the construction algorithm clearer but are not part of the gene tree
- An incidence matrix and a gene tree are **equivalent representations** of the same dataset
- To build a gene tree from some DNA data, we need to know for each polymorphic site which version of the site is the ancestral one (denoted by a 0 in the incidence matrix) and which one is the mutation (denoted by a 1 in the incidence matrix)
- This can often be found by considering the sequence of an **outgroup**
- Note that **a gene tree is not a coalescent tree**!

Example

Whitfield, Sulston and Goodfellow (1995) sequenced 18.3 Kbp from the Y chromosome of five humans and one common chimpanzee. Only three sites were polymorphic when comparing the five humans:

Subject	Site 1	Site 2	Site 3
European	G	C	A
Melanesian	G	T	G
Rondonian surui	G	C	G
Tsumkwe san	G	C	A
Mbuti pygmy	A	C	G
Chimpanzee	G	C	A

$$\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{pmatrix}$$

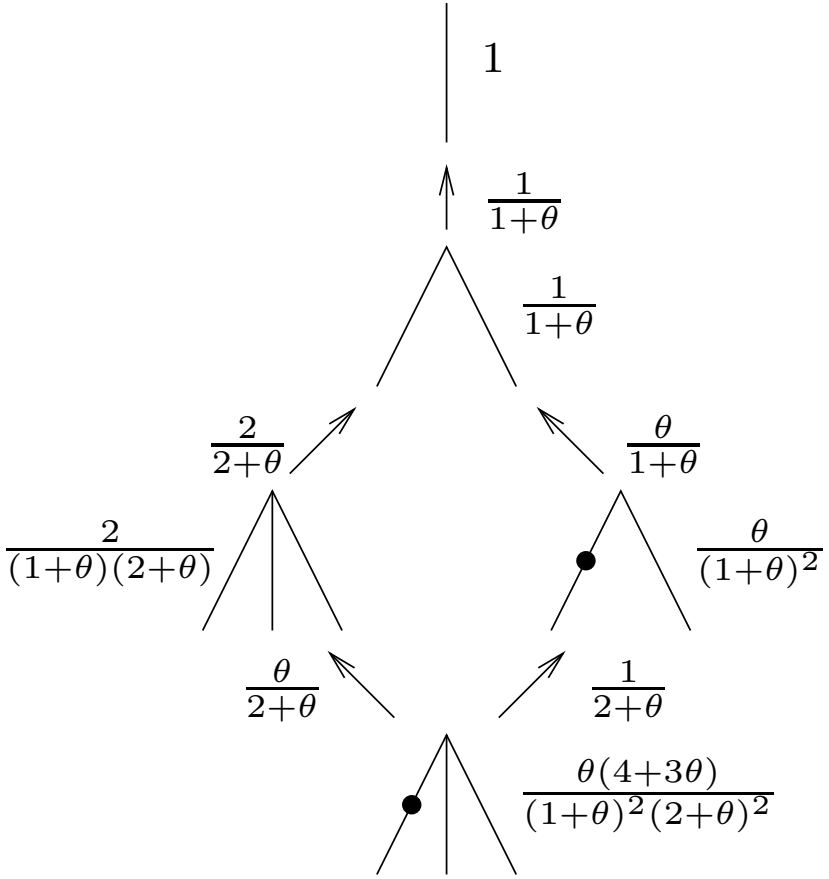


Probability of a gene tree

- **No known formula** in the infinite sites model to calculate the probability of data as the ESF does in the infinite alleles model.
- It is however possible to follow the same approach that we used in the proof of the ESF to establish a **recursion equation** which can be used to calculate the probability of a configuration for small datasets.
- This recursion was first established by Ethier and Griffiths (1987) and further investigated by Griffiths (1987, 1989) and Griffiths and Tavaré (1995).
- It is thus called **the Ethier-Griffiths-Tavaré recursion** (EGT)
- Here we don’t give the exact form of the EGT recursion which is a bit complex but simply show an example of application.

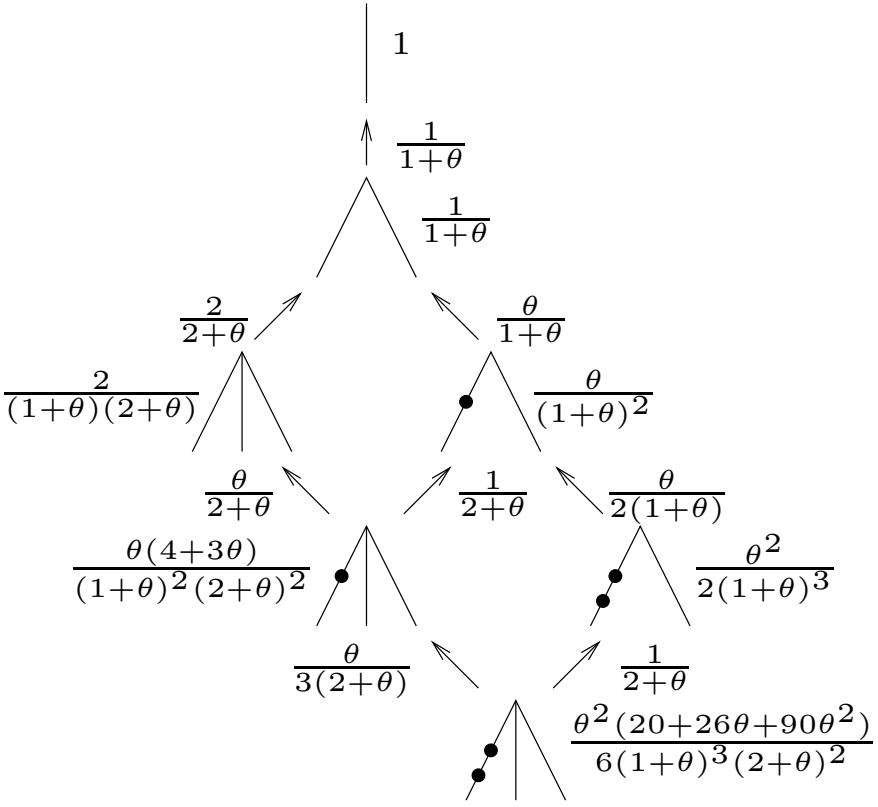
Example

When observing three sequences, what is the probability that two have no mutation and one has one mutation?



Example

When observing three sequences, what is the probability that two have no mutation and one has two mutations?



Summary

- In the **infinite sites model**, each mutation affects a new site
- Thus all mutations are observed, unlike in the infinite alleles model
- The **number of polymorphic sites** is equal to the number of mutations that occur ed
- A dataset can be represented as a **gene tree**
- The probability of a dataset can be computed using the **Ethier-Griffiths-Tavaré recursion**