Coalescent

C Neuhauser and S Tavaré

doi: 10.1006/rwgn.2001.1418

Neuhauser, C
Department of Ecology, Evolution and Behaviour, 100 Ecology Building, 1987 Upper Buford Circle, St. Paul, MN 55108, USA

Tavaré, S
Department of Mathematics, University of Southern California, Los Angeles, CA 90089, USA

Recent advances in molecular biology have made large-scale studies of molecular variability within populations a reality. Data from such studies are often obtained as random samples of DNA sequences, or as samples of single nucleotide polymorphisms. Because the individuals in the sample are related, these data are highly dependent; understanding the nature of this dependence is crucial for the analysis of the variability in the sample.

In contrast to data collected from pedigrees, the precise nature of the ancestral relationships among the DNA sequences in a random population sample is not known, and must be modeled. The coalescent, introduced by Kingman in 1982, describes one class of models for the genealogical relationships among a random sample of chromosomes.

The use of genealogical or coalescent methods is now central to the analysis of much genetic data. They allow for efficient simulation of the molecular structure of a sample of chromosomes; instead of simulating the entire population and then sampling from that, one only needs to keep track of the ancestors of the sample. Furthermore they provide a natural framework for estimation and inference about population parameters such as mutation rates and recombination rates, as well as about features of the ancestry of the sample or population.

The Ancestral Process

The Neutral Case

To describe the genealogy under a neutral model we assume that the population is haploid and of fixed size N individuals. Furthermore, we assume that the population evolves according to the discrete time Wright–Fisher model. In this model N descendants are chosen in each generation according to a multinomial distribution which reflects the gene frequencies in the previous generation. For instance, in the case of a single locus with two alleles A_1 and A_2 with respective frequencies x_1 and x_2 the probability that there are k descendants of type A_1 in the following generation is given by

$$\binom{N}{k} x_1^k x_2^{N-k}, \quad k = 0, 1, \ldots, N$$

if one ignores the possibility of mutation.

In the neutral case, demography and the mutation process can be separated. This allows one to determine the ancestral relationships in the sample without reference to the allelic types.

When the population size N is large compared to the sample size the genealogy of a sample of size n can be approximated by a continuous time Markov chain $A(t)$ in which time t is measured in units of N generations. The process starts from $A(0) = n$ and goes through the states $n, n-1, \ldots, 2, 1$. A value of $A(t) = j$ means that the sample had j distinct ancestors time t ago. The amount of time T_j for which there are j ancestors is exponentially distributed with mean $2/[j(j-1)]$, and these times are independent of one another. This Markov chain $A(t)$ is called the *coalescent process*.

Of interest is the time to the most recent common ancestor (MRCA) of the sample. This time is denoted by T_{MRCA}. It can be represented as the sum of the coalescence times T_j, that is,

$$T_{\mathrm{MRCA}} = T_n + T_{n-1} + \cdots + T_2$$

It follows that the expected time to the MRCA is $2(1 - 1/n)$. Thus in a large sample, the time to the MRCA is on average about $2N$ generations.

The genealogy can be visualized as a coalescing tree. A realization is shown in **Figure 1**. A tree that corresponds to a sample of size n has n tips and one root. The root is the location of the most recent common ancestor.

A characteristic of the neutral genealogy for fixed population size is that the last two branches dominate the height of the tree. This can be seen by comparing the expected coalescing time of two branches, ET_2, and the expected time to the most recent common ancestor, T_{MRCA}. The expected time until two ancestors coalesce is 1 which is more than half of the total expected time to the most recent common ancestor, regardless of the sample size.

Since under neutrality demography and the mutation process can be separated, to obtain a sample of size n, one can first construct its generalogy and then superimpose the mutation process on the genealogy. This provides an extremely efficient way to simulate

observations from complicated demographic and mutation scenarios.

We assume the simplest mutation process in which mutations occur independently to all genes with probability u_N per gene per generation. If time is scaled in units of N generations and if

$$\lim_{N\to\infty} 2Nu_N = \theta$$

then mutations occur along the branches of the coalescent process according to a Poisson process with rate $\theta/2$ independently in each branch of the coalescent.

The distribution of the total number of mutations in the sample since their most recent common ancestor follows readily. Given the total length T of the branches in the tree, which is

$$T = \sum_{j=2}^{n} jT_j,$$

the total number of mutations in the tree follows a Poisson distribution with mean $\theta T/2$.

The Selection Case

In the neutral case, the demography and the mutation process can be separated. This is reflected in the fact that the genealogy of a sample can be reconstructed without reference to the mutation process. Mutations can be superimposed on the genealogy. This separation of demography and mutation process no longer holds true when natural selection is incorporated into the model. Under selection reproductive success depends on the allelic type. This is reflected in the more complicated structure of the ancestral graph.

The simplest case of a population model with selection and mutation is a discrete time haploid Wright–Fisher model with two alleles A_1 and A_2 at one locus. Mutations from A_1 to A_2 or the reverse occur with probability u_N per gene per generation. Genes of type A_2 have a selective advantage with selection parameter s_N. That is, if $Y_1(k)$ denotes the number of gene of type A_1 at generation k, then

$$P[Y_1(k+1) = j | Y_1(k) = i] = \binom{N}{j} \psi_1^j (1-\psi_1)^{N-j}$$

with

$$\psi_1 = \frac{p(1-u_N) + (1-p)(1+s_N)u_N}{p + (1-p)(1+s_N)}$$

where $p = i/N$, the fraction of genes of type A_1 in generation k.

Again, when the population size is large, the genealogy of a sample of n genes can be approximated by a continuous time Markov process $G(t)$, $t \geq 0$. This limiting object is called the *ancestral selection graph*. Time t is measured in units of N generations and

$$\lim_{N\to\infty} 2Nu_N = \theta \quad \text{and} \quad \lim_{N\to\infty} 2Ns_N = \sigma$$

As in the neutral case, the genealogical process can be most easily explained when visualized as a graph. The ancestral graph has a coalescing/branching structure. An ancestral graph is shown in **Figure 2**. The ancestral graph is a stochastic process whose dynamics are as follows. If there are k branches in the graph, then a coalescence event occurs at rate $k(k-1)/2$, and a branching event occurs at rate $k\sigma/2$. Coalescing events correspond to the merging of two ancestral lines as in the neutral case. Branching events are a characteristic of genealogies under selection. They reflect the fact that the fitter type has a higher reproductive success than the less fit type. Following an ancestral line back on the ancestral graph, at a branching point the two branches coming out of a point constitute *possible* ancestral paths. The branch that branches off the straight branch in the graph is called the incoming branch, while the straight branch is called the original branch. If the ancestor on the incoming branch is of the fitter type, then the ancestral path follows the incoming branch; if not, it follows the original branch. Paths in the ancestral selection graph are thus *possible* ancestral paths. As long as $\sigma < \infty$, the size of the graph will eventually reach 1. The ancestor at this instant is called the *ultimate ancestor*. Which of the paths are contained in the embedded genealogy can be determined once the ultimate ancestor is found.

The type of the ultimate ancestor needs to be chosen according to the allele frequencies at the time of the ultimate ancestor. For instance, if the gene frequencies were in equilibrium at that time, the type of the ultimate ancestor would be chosen from the stationary distribution.

Mutation events can be treated as in the neutral case: mutation events are superimposed on the ancestral graph at rate $\theta/2$, independently in each branch.

Embedded in an ancestral recombination graph is the true genealogy of the sample, called the *embedded genealogy*. with its To find the embedded genealogy, one starts at the ultimate ancestor and follows the graph forward in time. At mutation events the type changes accordingly. At coalescing events, the two branches coming out of the coalescing point receive the same type as the branch entering the coalescing point. At branching points, if the incoming branch has the fitter allele, then the gene on the incoming branch

continues. Following these rules one eventually arrives at the present time and obtains a sample of size n. Going back up the graph one can then extract the embedded genealogy and identify, for instance, the most recent common ancestor. As shown in **Figure 3**, this may differ from the ultimate ancestor.

Robustness of the Genealogy

The coalescent is remarkably robust. It provides a good approximation for a large class of reproduction models when the population size N is large relative to the sample size n.

This class includes both discrete time models in which generations do not overlap and continuous time models in which generations overlap. One can also change the offspring distribution. For instance, if the variance of the offspring distribution is υ, then in the neutral case a change in the time scale of the coalescent occurs: The average time between coalescing events changes by a factor $1/\upsilon$. This implies that the time to the most recent common ancestor is shortened if the variance of the number of offspring is increased.

Furthermore, genealogies can be formulated for diploid populations. In the neutral case when mating is random (i.e., a panmictic population), diploidy simply means that the number of genes is doubled: if the population size is N, then the number of genes is $2N$. The genealogy in the diploid case is then the same as in the haploid case with N is replaced by $2N$. In the selective case when mating is random, the ancestral graph is more complicated. At branching points, three branches now come together. The additional branch is used to identify the type of the diploid parent. As in the haploid case it is possible to extract the embedded genealogy by following the paths in the ancestral graph.

Varying Population Size

It is straightforward to incorporate deterministically varying population size into the ancestral process. This only affects the coalescing rate and is therefore the same for both the neutral and the selective case.

If $N(t)$ denotes the population size t units in the past where t is measured in units of $N = N(0)$ generations and if $N(t)/N \to 1/\mu(t)$, then the coalescing rate is $k(k-1)\mu(t)/2$ if there are k branches present at time t.

The effect of a growing population can be quite dramatic. For instance, if the population has grown exponentially, i.e., $N(t) = e^{-\beta t}N$ for some $\beta > 0$, then $\mu(t) = e^{\beta t}$ and the coalescing rate is $k(k-1)e^{\beta t}/2$. The resulting graph is stretched near the present time and compressed in the past (i.e., near the root). The resulting graph resembles a star phylogeny in the neutral case.

Recombination

To describe the genealogy of two linked loci, L_1 and L_2, we assume that the population is of fixed size N and evolves according to the neutral Wright–Fisher model. Recombination occurs independently in each offspring. In each generation, with probability $1 - r$ each offspring independently inherits the genes at loci L_1 and L_2 from the same chromosome; with probability r the genes are inherited from different chromosomes (i.e., a recombination event occurred).

When the population is large and

$$\lim_{N\to\infty} 2Nr = \rho$$

the genealogy of a sample of size n can be approximated by a continuous time Markov chain $R(t)$, $t \geq 0$, where time t is measured in units of N generations. This Markov chain, known as the *ancestral recombination graph*, can be described as a graph that contains the lineages of each individual of the sample. Following a lineage backwards in time on this graph, recombination events occur at rate $\rho/2$. At such times, the lineage of the two loci L_1 and L_2 splits which results in a branching event. One branch follows the ancestry of one locus, the other branch follows the ancestry of the other locus. Common ancestry is again represented by the coalescing of branches. An example is given in **Figure 4**.

The dynamics of this recombination graph are given as follows. If there are k branches in the graph, then a coalescing event occurs at rate $\binom{k}{2}$, that is, each pair of branches coalesces at rate 1: a branching event in which a branch splits into two, occurs at rate $k\ \rho/2$, that is, each branch splits into two at rate $\rho/2$.

If one adopts the convention that branches that correspond to the L_1 locus are drawn to the left and branches that correspond to the L_2 locus are drawn to the right at branching points, then the ancestry of each locus can be traced separately by following the paths to the left for the L_1 locus and to the right for the L_2 locus at each branching point. It follows that the ancestry of each locus is given by the neutral coalescent process and each subtree has its own most recent common ancestor. These marginal coalescent trees are of course not independent of one another.

The ancestral graph can be adapted to describe multiple loci by keeping track of where the breakpoints occur at each recombination event. Just as earlier, mutations can be superimposed on the ancestral

recombination graph at rate $\theta/2$, independently in each branch.

Migration and Subdivision

The assumption of panmixia can be replaced by the assumption that the population is geographically structured. The simplest case is that of a subdivided population in which the popularion consists of a finite number of islands, each populated by a subpopulation. The size of the subpopulation on island i is denoted by N_i for $i = 1, 2, \ldots, K$, where K is the total number of islands. Reproduction on each island follows the Wright–Fisher model (possibly with selection). Each generation, a proportion m_{ij} of the offspring on island i migrates to island j, regardless of their genotype. A simplifying assumption is to stipulate that the sizes of subpopulations are fixed, that is, immigration balances emigration at all times.

When the sizes of all subpopulations are sufficiently large, the genealogy of a sample of size n can be approximated by a continuous time Markov chain $S(t)$, $t \geq 0$, where time t is measured in units of $N = \sum_{i=1}^{K} N_i$ generations. This process is called the *structured coalescent*. In addition to the coalescent process in each of the islands, each branch in island i, $i = 1, 2, \ldots, K$, "migrates" to island j at rate $\mu_{ij}/2$ where

$$\lim_{N\to\infty} 2N\frac{N_i}{N_j}m_{ij} = \mu_{ij}$$

The effect of population subdivision compared to the panmictic case is a compression of the coalescent near the tips of the tree due to the smaller sizes of the subpopulations. However, further back in the past, the branches are extended provided the migration rate is small enough since lineages have to be on the same island in order to coalesce.

There are cases where these effects balance each other and the mean coalescing time for a pair of genes with population subdivision is identical to the panmictic case. However, the coalescing time in the subdivided population shows much greater variance than in the panmictic case.

Strong Selection

Under selection, demography and mutation become inseparable which results in a more complicated ancestral process. However, if selection is sufficiently strong, one can again separate demography and mutation, at least approximately. The embedded genealogy then becomes approximately a simple time change of Kingman's neutral coalescent. The reason for this is that under strong selection the population dynamics are on a much faster time scale than coalescing events.

In cases where this separation of time-scale occurs, the ancestral process can be modeled as a change in the effective population size. In particular, this says that not only are the expected times between coalescing events a time change relative to the neutral case but the distribution of the coalescing events is the same as in the neutral case except for the time-scale.

Strong selection can often be modeled as a subdivided population where the subpopulations correspond to the different alleles. Migration between subpopulations is then governed by the mutation process.

Other Coalescents

The structure of the coalescent has been identified for a wide variety of other phenomena, such as nonrandom mating (e.g., selfing), different sexes, age structure, and so on. We have assumed in our exposition that mutation, recombination, and selection rates are of the order of the reciprocal of the population size. In cases where this is not true, other behavior for the genealogy is possible; discrete time branching processes arise in this context.

Inference

An important use of coalescents arises when using random population samples to estimate population parameters such as ρ, θ, and σ. A number of approaches have been proposed for this purpose, including those based on the behavior of summary statistics (for example, the number of segregating sites observed in a sample of DNA sequences is often used to estimate θ). Full likelihood methods and Bayesian approaches are currently of great interest, particularly as they provide an inferential framework for mapping disease genes by linkage disequilibrium mapping, and by haplotype sharing. Importance sampling and Markov chain Monte Carlo approaches have proved useful in this context.

Further Reading

Donnelly P and Tavaré S (1995) Coalescents and genalogical structure under neutrality. *Annual Review of Genetics* 29: 410–421.

Donnelly P and Tavaré S (eds.) (1997) *Progress in Population Genetics and Human Evolution, IMA Proceedings*, vol. 87 New York: Springer-Verlag.

Fu Y-X and Li W-H (1999) Coalescing into the 21st century: an overview and prospects of coalescent theory. *Theoretical Population Biology* 56: 1–10.

Griffiths RC (1991) The two-locus ancestral graph. In: *Selected Proceedings of the Symposium on Applied Probability, Sheffield*, 1989, vol.18 of *IMS Lecture Notes-Monograph Series*, pp. 100–117. Institute of Mathematical Statistics.

Hudson RR (1983) Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology* 23: 183–201.

Hudson RR (1991) Gene genealogies and the coalescent process. In: Futuyma D and Antonovics J (eds) *Oxford Surveys in Evolutionary Biology*, vol. 7, pp. 1–44. New York: Oxford University Press.

Kingman JFC (1982a) On the genealogy of large populations. *Journal of Applied Probability* 19A: 27–43.

Kingman JFC (1982b) The coalescent. *Stochastic Processes and Applications* 13: 235–248.

Neuhauser C and Krone SM (1997) The genealogy of samples in models with selection. *Genetics* 145: 519–534.

Nordborg M (2001) Coalescent theory. In: Balding DJ, Cannings C and Bishop M (eds) *Handbook of Statistical Genetics*, pp. 000–000. Chichester, UK: Wiley.

Nordborg M and Tavaré S (2001) Linkage disequilibrium, haplotype sharing, and the coalescent. *Trends in Genetics* 000–000.

Stephens M and Donnelly P (2000) Inference in molecular population genetics. *Journal of the Royal Statistical Society B* 62: 605–635.

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105: 437–460.

***See also:* 0803 (Maximum Likelihood), 0957 (Parsimony), 1479 (Trees), 0532 (Genetic Distance), 0995 (Phylogeny)**

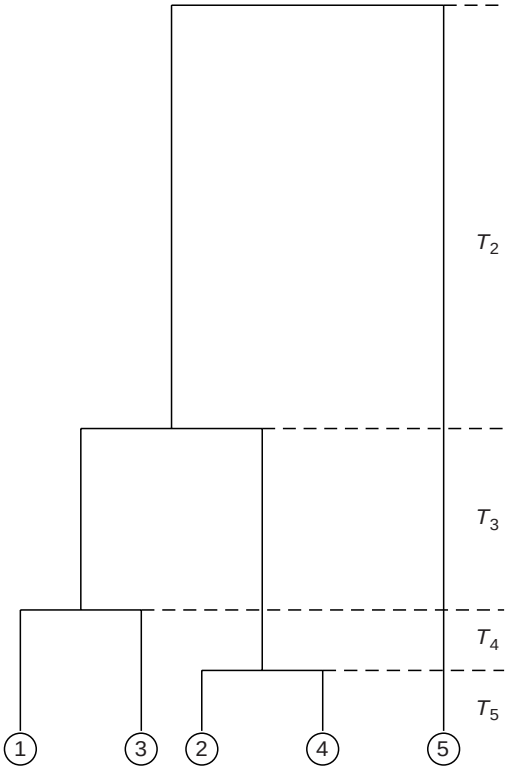

Figure 1 Coalescent tree of sample of five individuals.

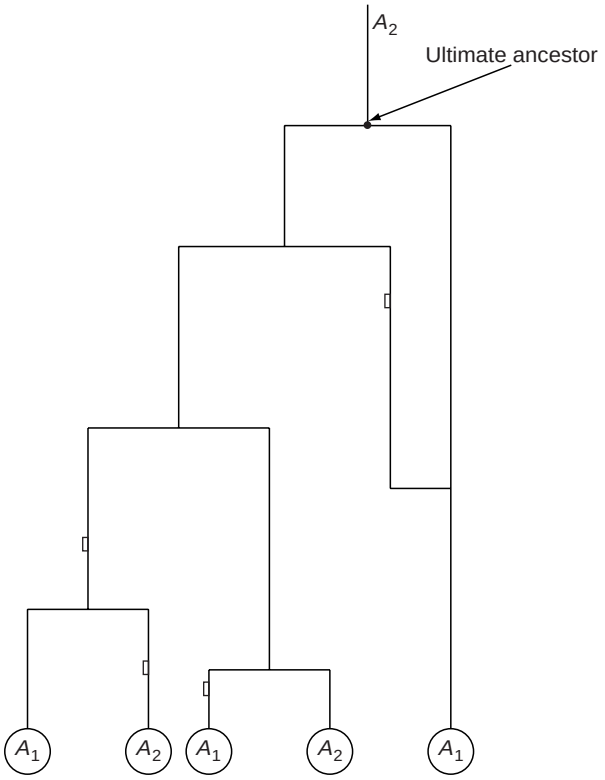

Figure 2 Ancestral selection graph for a sample of five individuals. Mutations denoted by [].

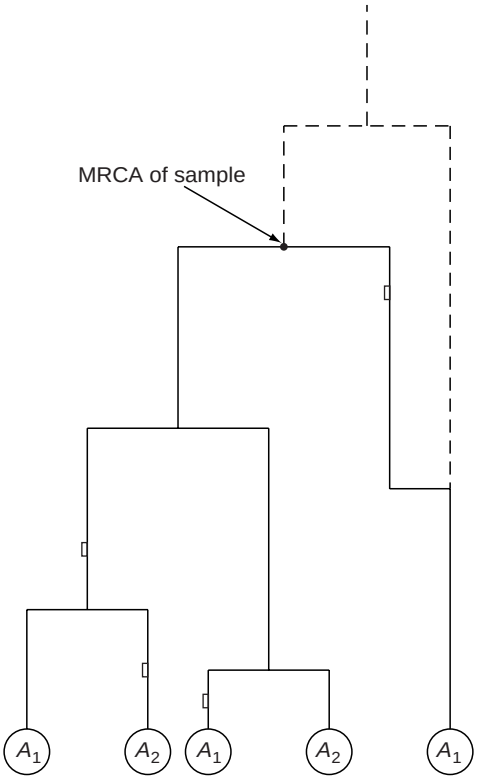

Figure 3 Embedded genealogy from Figure 2. Mutations denoted by [].

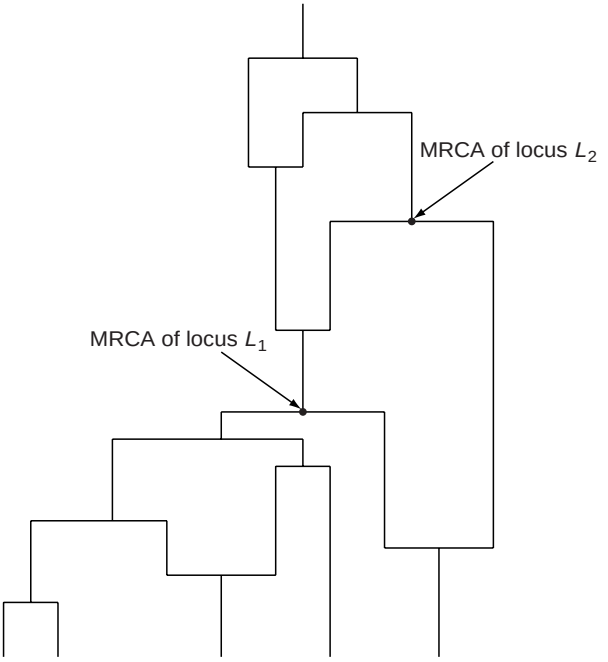

Figure 4 Two-locus ancestral recombination graph for sample of five individuals.