An introduction to coalescent theory

Nicolas Lartillot

May 29, 2012
Inferring population history from haplotype data

- a set of $n$ haplotypes randomly sampled from a population
- sequences of length $L$, known mutation rate $\mu$
- what can we say about
  - population size ($N$) and structure?
  - demographic history?
  - selection?
Approach

- define a model of demography and reproduction (Wright-Fisher)
- induces a law on gene genealogies (Kingman’s coalescent)
- then define a model of DNA sequence mutations
- explain variation in gene sample based on combination of mutation and coalescent models.

Applications

- estimating parameters (population size, mutation rate)
- testing hypotheses (e.g. deviation from neutrality)
- building blocks for more sophisticated models (course no 2)
The Wright-Fisher model

Assumptions

- panmictic population
- constant population size
- neutral
The Wright-Fisher model

- each offspring 'chooses' parent uniformly among $2N$ individuals of previous generation
- distribution of number of offspring: $\text{Binomial}(2N, 1/2N)$
Genealogy of a sample

- $n$ individuals taken at random (here $n = 3$)
- Age of their ancestor?
- Typical shape of the genealogy?
coalescence of $n = 2$ genes

- prob. of coalescence in previous generation $1/(2N)$
- average coalescence time for 2 individuals: $\bar{T} = 2N$. 
Relation between genetic diversity and coalescence time \((n = 2)\)

- time since last common ancestor: \(T\) generations
- sequences of length \(L\), known mutation rate \(\mu\)
- mean fraction of sites differing between 2 individuals: \(\pi = 2\mu T\)
coalescence of $n = 2$ genes

with mutation

- mutations at rate $\mu$ per base pair per generation
- average diversity: $\pi = 2\bar{T}\mu = 2.2N.\mu = 4N\mu = \theta$.
- $\theta$: scaled mutation rate (N and $\mu$ are confounded)
- yields an estimate of $N$ if $\mu$ is known and $\pi$ is observed
Tajima’s estimator

$n = 4$ observed DNA sequences

\begin{align*}
1 & \quad A \quad C \quad C \quad A \quad A \quad A \quad A \quad G \\
2 & \quad A \quad C \quad C \quad A \quad G \quad T \quad A \quad G \\
3 & \quad A \quad C \quad T \quad G \quad C \quad A \quad T \quad G \\
4 & \quad A \quad C \quad T \quad G \quad G \quad T \quad A \quad C \\
\end{align*}

$\pi_{ij}$: fraction of polymorphic sites between haplotypes $i$ and $j$

$$\hat{\pi} = \frac{2}{n(n - 1)} \sum_{i < j} \pi_{ij}$$
Effective population size of humans

Human-chimp divergence
- SND (single nucleotide differences): $\approx 2\%$
- divergence time: $\approx 6Ma$
- thus, mutation rate: $\approx 3 \times 10^{-8}$

Human polymorphism
- heterozygosity: $\pi = 0.001$ (1 every 1000 bp)
- SNP (single nucleotide polymorphisms): 1 every 100 to 300 bp

\[
\pi = 4N\mu \\
N = \frac{\pi}{4/\mu} \approx 10000
\]

- effective population size $\lt$ census population size
Effective population size

Genetic aspects

- autosomal: $2N$
- X chromosome: $3/2 \cdot N$
- mitochondrial, Y chromosome: $N$

Demographic aspects

- $N$: harmonic mean of census size over short-term fluctuations
- frequent bottlenecks: low $N$
- reproductive variance (species with male dominance have low $N$)
- population structure (e.g. a parasite has $N$ of its host)

Linkage and selection

- selection at linked loci reduce $N$ at neutral loci
- purifying selection: background selection
- positive selection: selective sweeps
Nucleotide diversity across life forms

Effective population size x Nucleotide mutation rate ($N_e u$)

Prokaryotes
- Prochlorococcus
- Tetrahymena thermophila
- Salmonella enterica
- Legionella pneumophila
- Helicobacter pylori
- Neisseria meningitidis
- Escherichia coli
- Vibrio cholerae
- Enterococcus faecium
- Cryptococcus neoformans
- Campylobacter jejuni
- Cryptosporidium parvum
- Saccharomyces cerevisiae
- Chlamydomonas reinhardtii
- Dictyostelium discoideum
- Neurospora crassa
- Streptococcus pyogenes
- Pseudomonas aeruginosa
- Giardia lamblia
- Toxoplasma gondii
- Trypanosoma cruzi
- Leishmania donovani
- Drosophila spp.
- Encephalitozoon cuniculi
- Artemia franciscana
- Zea mays
- Caenorhabditis spp.
- Arabidopsis thaliana
- Ciona intestinalis
- Silene spp.
- Crassostrea virginica
- Anopheles spp.
- Strongylcentrotus franciscanus
- Pinus sylvestris
- Fugu rubripes
- Hordeum vulgare
- Plasmodium falciparum
- Oryza sativa
- Ficedula spp.
- Oncorhynchus tshawytscha
- Pan troglodytes
- Homo sapiens
- Mus musculus

Unicellular eukaryotes

Invertebrates
- Vascular plants

Vertebrates
Effective population sizes across life forms

Mutation rates (per generation)

- human: $\approx 10^{-8}$
- fly, nematode: $\approx 10^{-9}$
- unicellular eukaryotes and prokaryotes: $\approx 10^{-10}$

Effective population sizes

- human, large vertebrates: $10^4$
- small vertebrates: $10^5$
- invertebrates, terrestrial plants: $10^6$
- unicellular eukaryotes: $10^7$
- prokaryotes: $> 10^8$
Population size and evolutionary genomics

Effective size and selection

- random drift proportional to $1/N$
- selection efficient only if $s \gg 1/N$

Evolutionary genomics

- small $N$: random drift dominates molecular evolution in humans
- many features selected in fly/yeast/E.coli not selected in humans
- genome structure influenced by population genetics parameters

Lynch and Conery, Science 2003; 302:1401
Distribution of age of ancestor

- prob. of coalescence in previous generation $\frac{1}{2N}$
- prob. of coalescence in 2 generations $(1 - \frac{1}{2N})(\frac{1}{2N})$
- prob. of coalescence in $t$ generations $(1 - \frac{1}{2N})^{t-1}\left(\frac{1}{2N}\right)$
- $t$ has a geometric distribution
Exponential distribution

$u = t/2N, \quad p(u) = e^{-u}$

- age of ancestor of 2 individuals has geometric distribution
- for $n << N$, approx. an exponential distribution
- mean of $t_2$ is $2N$, (std dev of $t_2$ is $2N$)
- rescaling: $u_2 = t_2/(2N)$ has mean 1 and stdev 1
Generalization for $n > 2$

- make Wright-Fisher simulations (pop. size $2N$)
- for each simulation, take $n$ chromosomes at final time (present)
- trace back their genealogy
- measure $t_j$ (in generations) and set $u_j = t_j/2N$ (rescaling)
- distribution of $t_j$ and $u_j$ over simulations?

rate of coalescence

\[
\begin{align*}
    r_2 &= 1/2N \\
    r_j &= \binom{j}{2} \frac{1}{2N} = \frac{j(j-1)}{4N}
\end{align*}
\]
Generalization for $n > 2$

- make Wright-Fisher simulations (pop. size $2N$)
- for each simulation, take $n$ chromosomes at final time (present)
- trace back their genealogy
- measure $t_j$ (in generations) and set $u_j = t_j/2N$ (rescaling)
- distribution of $t_j$ and $u_j$ over simulations?

![Mean coalescence times](image)

- $\bar{t}_2 \approx 2N$
- $\bar{t}_j \approx \frac{4N}{j(j-1)}$, $j = 2..n$
Generalization for $n > 2$

- make Wright-Fisher simulations (pop. size $2N$)
- for each simulation, take $n$ chromosomes at final time (present)
- trace back their genealogy
- measure $t_j$ (in generations) and set $u_j = t_j / 2N$ (rescaling)
- distribution of $t_j$ and $u_j$ over simulations?

### Mean coalescence times

- $\bar{u}_2 \approx 1$
- $\bar{u}_j \approx \frac{2}{j(j-1)}$, $j = 2 \ldots n$
Generalization for $n > 2$

- make Wright-Fisher simulations (pop. size $2N$)
- for each simulation, take $n$ chromosomes at final time (present)
- trace back their genealogy
- measure $t_j$ (in generations) and set $u_j = t_j / 2N$ (rescaling)
- distribution of $t_j$ and $u_j$ over simulations?

\[
\begin{align*}
t_j & \sim \text{Exp} \left( \text{mean} = \frac{4N}{j(j - 1)} \right) \\
u_j & \sim \text{Exp} \left( \text{mean} = \frac{2}{j(j - 1)} \right)
\end{align*}
\]
Drawing from the coalescent

Algorithm

for \( j = n \ldots 2 \):

- draw \( u_j \sim \text{Exp} \left( \text{mean} = \frac{j(j-1)}{2} \right) \)
- join 2 of the \( j \) remaining lineages taken at random
Drawing from the coalescent

Forward versus backward simulation
- forward: Wright Fisher simulation + backtracking of ancestors
- backward: Kingman’s coalescent: drawing exponential variables
- equivalence ($n \ll N$), but
- Kingman’s approach more efficient (in $n$ instead of $N^2$)
Drawing from the coalescent

- large variability of deep branches
- high uncertainty on population size estimate based on one locus
- suggests approaches averaging over several independent loci
What is coalescent theory useful for?

Theory
- obtaining insights about patterns in sequence variation
- deriving theoretical expectations (e.g. age of sample’s last common ancestor)

Simulations
- null distribution for hypothesis testing
- detecting departures from neutrality (selection)

Parameter estimation
- estimating $\theta = 4Nu$ based on observed polymorphism
- estimating demographic scenarios (see course 2)
Mean age of most recent common ancestor (MRCA)

\[ T_n = u_n + u_{n-1} + \ldots + u_2 \]

\[ E[T_n] = 2(1 - 1/n) \]

- expected MRCA age reaches a limit (4N generations) for large \( n \)
- intra-specific variation gives access to relatively shallow past
- in contrast to interspecific divergence (human chimp: 6 Myrs)
Age of most recent common ancestor

- mitochondrial: 200 000 years (Soares et al, 2009, Am J Human Genet 84:740)
- Y chromosome: 55 000 years (Thomson et al, 2000, PNAS, 97:7360)
- nuclear genome: variation along genome
Genealogies and recombination

Here we introduce the most popular population genetics model: the coalescent. We begin by introducing the simplest form, in which there is no recombination, and then discuss the version that applies in a more realistic setting.

Coalescent without recombination
Panels a–c illustrate the intuition that underlies the coalescent using a population of DNA fragments that are evolving according to a Wright–Fisher model — that is, in the absence of recombination, in a population of constant size.

Panel a shows a schematic of an evolving population. In this simplified representation of evolution, each row corresponds to a single generation, and each blue circle denotes a fragment in that generation. Generations are replaced in their entirety by their offspring, with arrows running from the parental fragment to the offspring fragment. The present day is represented by the bottom row, with each higher row representing one generation further back into the past.

Panel b indicates the ancestry of a sample from the present day. In this example, six fragments, indicated in red, are sampled from the current generation. The ancestry of this sample is then traced back in time (that is, up the page), and is indicated in red.

Panel c highlights one of the key features of the coalescent: all information outside the ancestry of the sample of interest can be ignored. The coalescent provides a mathematical description of the ancestry of the sample. As we move back in time, the number of lines of ancestry decreases until, ultimately, a single line remains. The most recent fragment from which the entire sample is descended is known as the 'most recent common ancestor' (MRCA), whereas the time at which the MRCA appears is known as the 'time to the most recent common ancestor' (TMRCA).

Coalescent with recombination
The coalescent with recombination is illustrated in panel d. In such settings, lines bifurcate, as well as coalesce (join), as we move back in time. Here we show the genealogy for three copies of a fragment. By tracing the lineages back in time, we observe the following events: in event 1 the green lineage undergoes recombination and splits into two lineages, which are then traced separately; in event 2 one of the resulting green lineages coalesces with the red lineage, creating a segment that is partially ancestral to both green and red, and partially ancestral to red only; in event 3 the blue lineage coalesces with the lineage created by event 2, creating a segment that is partially ancestral to blue and red, and partially ancestral to all three colours; in event 4 the other part of the green lineage coalesces with the lineage created by event 3, creating a segment that is ancestral to all three colours in its entirety. As the inset shows, the recombination event induces different genealogical trees on either side of the break.

Coalescent methods have been reviewed extensively, and there are now book-length treatments to which the reader is referred for further details.

Panel d is modified with permission from REF. 89 © (2002) Elsevier.

Marjoram and Tavaré, 2006, Nat Rev Genet, 7:759
The basic idea underlying the coalescent is that, in the absence of selection, sampled lineages can be viewed as randomly 'picking' their parents, as we go back in time (FIG. 4). Whenever two lineages pick the same parent, their lineages coalesce. Eventually, all lineages coalesce into a single lineage, the MRCA of the sample. The rate at which lineages coalesce depends on how many lineages are picking their parents (the more lineages, the faster the rate) and on the size of the population (the more parents to choose from, the slower the rate).

Mutation and recombination. We now need a population-genetics model that incorporates these principles and that allows us to construct and analyse random genealogies. The coalescent has become the standard model for this purpose. This choice is not arbitrary, as the coalescent is a natural extension of classical population-genetics theory and models. It was discovered independently by several authors in the early 1980s, although the definitive treatment is due to Kingman.

Rosenberg and Nordborg, 2002, Nat Rev Genet, 3:380
Total length of the genealogy

\[ L_n = \sum_{j=2}^{n} j u_j \]

\[ E[L_n] = \sum_{j=2}^{n} j \frac{2}{j(j-1)} \]

\[ = 2 \sum_{j=2}^{n} \frac{1}{j-1} \]

for large \( n \)

\[ E[L_n] \sim 2 \ln n \]

(slow increase)
Estimating $\theta = 4N\mu$: Watterson’s estimator

- $S_n$: number of sites segregating in the sample
- low mutation rate: $S_n =$ total # mutations along genealogy

$$\hat{\theta} = \frac{2S_n}{E[L_n]}$$

$$E[S_n] = 2N\mu E[L_n] = \theta E[L_n]/2$$

$$L_n = \sum_{j=2}^{n} j u_j$$

$$E[L_n] = 2 \sum_{j=2}^{n} \frac{1}{(j-1)}$$
Estimating $\theta = 4N\mu$: Tajima versus Watterson

Tajima’s estimator of scaled mutation rate

- $\pi_{ij}$: fraction of polymorphic sites between haplotypes $i$ and $j$

$$\hat{\pi} = \frac{2}{n(n-1)} \sum_{i<j} \pi_{ij}$$

Watterson’s estimator

- $S_n$: number of sites segregating in the sample
- $E[L_n]$: mean total length of genealogy

$$\hat{\theta} = \frac{2S_n}{E[L_n]}$$
Variance of the two estimators

- Tajima’s estimator is not consistent
- Watterson’s estimator consistent but not optimal
- maximum likelihood (see later) optimal and more general

Felsenstein 1992
Demography and population structure

- changes in population size induce changes in rate of coalescence
- at time $t$, rate of coalescence of $j$ lineages is $j(j - 1)/4N(t)$
- increasing population: comparatively higher rates in distant past
- decreasing population: comparatively higher rates near present
Tajima’s and Watterson’s estimates respond differently to changes in $N$

- increasing population: $d = \hat{\pi} - \hat{\theta} < 0$
- decreasing population: $d = \hat{\pi} - \hat{\theta} > 0$
- Tajima’s $D = d / \hat{\mathbb{V}}(d)$
Hypothesis testing using Tajima’s D

Principle

- estimate $\hat{\pi}$ and $\hat{\theta}$, compute $D$
- simulate genealogies and distribute mutations over it with rate $\hat{\theta}$
- on each replicate, estimate $\hat{\pi}$ and $\hat{\theta}$, compute $D$: null distribution

Scope and limits

- significant deviation: departure from any assumption
- demography ($D < 0$: population increase)
- selection ($D < 0$: directional selection, $D > 0$ balancing selection)
- panmixia (but $D$ is more robust to this)
Selective sweep: the process by which a new advantageous mutation eliminates or reduces variation in linked neutral sites as it increases in frequency in the population in explaining the pattern of variability within and between species (39, 59).

Much of the theoretical literature in population genetics over the past 50 years has focused on developing and analyzing models that generalize the previously mentioned basic di-allelic models to models where more than two alleles may be segregating, where multiple mutations may arise and interact—possibly in the presence of recombination, where the environment may be changing through time, and where random genetic drift may be acting in populations subject to various demographic forces (25, 39). From theory alone we have gained many valuable insights, including the fact that the efficacy of selection depends not only on the selection coefficient, but primarily on the product of the selection coefficient and the effective population size. An increased effect of selection may be due to either an increased population size or a larger selection coefficient. Among other important findings is that balancing selection may occur for many reasons other than overdominance, (e.g., fluctuating environmental conditions) and could therefore, potentially, be quite common (38, 39). However, the efficacy of selection will tend to be reduced when multiple selected alleles are segregating simultaneously in the genome. The mutations will tend to interfere with each other and reduce the local effective population size (8, 29, 40, 57). Many population geneticists used to believe that the number of selective deaths required to maintain large amounts of selection would have to be so large that selection would probably play a very small role in shaping genetic variation (43, 60, 61). These types of arguments, known as genetic load arguments, were instrumental in the development of the neutral theory. However, the amount of selection that a genome can permit depends on the way mutations interact in their effect on organismal fitness and on several other critical model assumptions (25, 62, 71, 107). Population genetic theory does not exclude the possibility that selection is very pervasive and cannot alone determine the relative importance and modality of selection in the absence of data from real living organisms (25, 39).

Much excitement currently exists in the population genetics communities over the fact that many predictions generated from the theory may now be tested in the context of the large genomic data sets. In particular, we should be able to detect the molecular signatures of new, strongly selected advantageous mutations that have recently become fixed (reached a frequency of one in the population). As these mutations increase in frequency, they tend to reduce variation in the neighboring region where neutral variants are segregating (13, 51, 52, 68). This process, by which a selected mutation reduces variability in linked sites as it goes to fixation, is known as a selective sweep (Figure 1). The hope is that by analysis of large comparative genomic data sets and large SNP data sets we will be able to determine how and where both positive and negative selection directional selection like population increase (at selected locus) locally in genome, looks like demographic expansion recombination progressively dissipates linkage with nearby neutral polymorphisms

Extensions to Kingman’s coalescent

- with demographic variation (time-dependent $N(t)$)
- with population structure (demes with migration between demes)
- with recombination (ancestral recombination graphs)
  - important tool for estimating recombination rates along genomes
- with selection (ancestral selection graphs)
Genealogies and recombination

Here we introduce the most popular population genetics model: the coalescent. We begin by introducing the simplest form, in which there is no recombination, and then discuss the version that applies in a more realistic setting.

Coalescent without recombination

Panels a–c illustrate the intuition that underlies the coalescent using a population of DNA fragments that are evolving according to a Wright–Fisher model — that is, in the absence of recombination, in a population of constant size.

Panel a shows a schematic of an evolving population. In this simplified representation of evolution, each row corresponds to a single generation, and each blue circle denotes a fragment in that generation. Generations are replaced in their entirety by their offspring, with arrows running from the parental fragment to the offspring fragment. The present day is represented by the bottom row, with each higher row representing one generation further back into the past.

Panel b indicates the ancestry of a sample from the present day. In this example, six fragments, indicated in red, are sampled from the current generation. The ancestry of this sample is then traced back in time (that is, up the page), and is indicated in red.

Panel c highlights one of the key features of the coalescent: all information outside the ancestry of the sample of interest can be ignored. The coalescent provides a mathematical description of the ancestry of the sample. As we move back in time, the number of lines of ancestry decreases until, ultimately, a single line remains. The most recent fragment from which the entire sample is descended is known as the 'most recent common ancestor' (MRCA), whereas the time at which the MRCA appears is known as the 'time to the most recent common ancestor' (TMRCA).

Coalescent with recombination

The coalescent with recombination is illustrated in panel d. In such settings, lines bifurcate, as well as coalesce (join), as we move back in time. Here we show the genealogy for three copies of a fragment. By tracing the lineages back in time, we observe the following events: in event 1 the green lineage undergoes recombination and splits into two lineages, which are then traced separately; in event 2 one of the resulting green lineages coalesces with the red lineage, creating a segment that is partially ancestral to both green and red, and partially ancestral to red only; in event 3 the blue lineage coalesces with the lineage created by event 2, creating a segment that is partially ancestral to blue and red, and partially ancestral to all three colours; in event 4 the other part of the green lineage coalesces with the lineage created by event 3, creating a segment that is ancestral to all three colours in its entirety. As the inset shows, the recombination event induces different genealogical trees on either side of the break.

Coalescent methods have been reviewed extensively, and there are now book-length treatments to which the reader is referred for further details.

Panel d is modified with permission from REF. 89 © (2002) Elsevier.

Marjoram and Tavaré, 2006, Nat Rev Genet, 7:759
Ancestral recombination graph: 2 loci

from Awadalla (McVean, Awadalla and Fearnhead, Genetics, 160:1231)

- scaled recombination rate $\rho = 4Nr$
- coalescence at rate $j(j - 1)/2$
- recombination at rate $j\rho/2$
Ancestral recombination graph: continuous segment of loci

Hein, Shierup and Wiuf, 2005

- scaled recombination rate (for whole segment) $\rho = 4Nr$
- coalescence at rate $j(j - 1)/2$
- recombination at rate $j\rho/2$
Lineage sorting

Hobolth et al, PLoS Genetics, 2007, 3 p.e7
Lineage sorting: structured coalescent

Probability of locus genealogy

\[ p(\text{HC1}) = 1 - e^{-2\tau_2/N_{HC}} \]
\[ p(\text{HC2}) = p(\text{HG}) = p(\text{CG}) = \frac{1}{3} e^{-2\tau_2/N_{HC}} \]
\[ p(\text{HC}) = p(\text{HC1}) + p(\text{HC2}) \]
Estimating ancestral population size

Tree mismatch approach (Nei 1987)

- for each locus, reconstruct most likely tree
- count proportions of trees = HC, HG or CG
- solve equation (last slide) for $\frac{\tau_2}{N_{HC}}$
- assuming $\tau_2 = 1.6$ Myrs, this yields $N_{HC} = 100,000 \pm 50,000$.

Problems

- bias due to stochastic tree reconstruction errors
- even under no lineage sorting, trees might differ due to finite alignment size
- results in an inflated estimate for $N_{HC}$
- need to use probabilistic models to improve on this estimate
Summary and conclusions

Summary

- rate of coalescence of $j$ lineages is $j(j - 1)/4N$
- depth of genealogy reflects population size
- shape of genealogy reflects demographic history
- Kingman’s coalescent: simple and powerful model for
  - understanding population genetics
  - estimating parameters
  - testing models

From there

- coalescent at the core of probabilistic models for statistical inference
- represents the natural law for integrating over unknown genealogies