1

TREES OF GENES IN POPULATIONS

Joseph Felsenstein

Abstract

Trees of ancestry of copies of genes form in populations, as a result of the randomness of birth, death, and Mendelian reproduction. Considering them allows us to think about evolution within and between populations, to make the connection between phylogenies and population genetic analyses. These trees, known as coalescents, are essential to developing methods for making inferences about populations. This chapter reviews coalescents and the inference methods based on them. The review concentrates on the population processes, and also briefly treats the inference methods, concentrating on those that attempt a likelihood or Bayesian treatment.

1.1 Introduction

Molecular evolution represents phylogenies as branching diagrams composed of thin lines. At the tip we often find one molecular sequence, sometimes described as “the yeast sequence” or “the mouse sequence”. It is as if we were viewing the evolutionary tree from a great distance, so that each branch appears thin. If each of these thin lines truly contained only one copy of this gene’s sequence, we would have a species that consisted only of a single individual, and a haploid one at that. But the lines are not lineages of single copies. Coming closer to them, we find that in reality the lines are thick – they are whole species, consisting of multiple populations, each of many individuals. To understand what molecular evolution looks like when we consider whole populations, we have to consider population-genetic phenomena in addition to the usual models of molecular evolution. The two fields of molecular evolution and population genetics (or evolutionary genetics) have grown up largely separately. However, they are connected, and with the availability of large population samples of sequences, their connections are increasing. We are well into a Great Encounter – the mathematics and statistics of population processes are becoming more and more important to molecular evolution, and multispecies comparisons are becoming more and more important to evolutionary genetics.

To explain how population-genetic models relate to molecular evolution between species, we have to start within species and model the ancestry of a population sample of n copies of a gene drawn from a single random-mating population. This ancestry is itself a tree, but not one whose forks are speciations. Instead they are simply events in which one parent copy gives rise to two or more offspring copies, a routine occurrence. The resulting trees have come to be called *coalescents*. They are sometimes called "gene trees", but this is ambiguous terminology, as that same phrase is also used for trees of descent of genetic loci by gene duplication, an entirely different phenomenon.

The most standard model of theoretical population genetics is the Wright-Fisher model. In it, each of the $2N$ copies of a gene in a diploid population of constant size N in effect chooses its parent copy from among the $2N$ parent copies available. These choices are independent. Thus for two copies in a population, there is a chance $1/(2N)$ that they came from the same copy in the previous generation. If they do not, the process occurs again when we go back one more generation. In effect, we toss a coin for each generation back, with the probability of Heads equal to $1/(2N)$. The time to the first Heads is drawn from a geometric distribution with that probability of Heads. This much was known to Sewall Wright and R. A. Fisher in the early 1930s.

In 1982, the eminent probabilist J.F.C. Kingman, who has had a lifelong interest in population genetics, asked what the process of ancestry would look like if we traced back from a sample of n copies in a large population of N individuals. He defined an excellent approximation which he called the *n-coalescent* [29, 30]. In it, one goes back in continuous time rather than in discrete generations. The ancestry of the n copies remains distinct for a time T_n generations, where T_n is drawn from an exponential distribution:

$$T_n \;\sim\; \mathrm{Exp}\left[4N/(n(n-1))\right]. \tag{1.1}$$

At that time two lineages chosen at random join, so that there are now $n-1$ lineages. The process then starts again, going back farther in time, but with the value of n decremented, as an independent draw from the same distribution with that smaller value of n. This continues until there are only two lineages, whose common ancestor is drawn by this process with $n = 2$.

Note that in the Wright-Fisher model the ancestry of copies of a gene can be discussed without considering whether or not the copies have the same or different DNA sequences. For the moment, there is assumed to be no natural selection. The copies reproduce in ways that do not depend on their DNA sequences.

This is an approximation to the genealogy implied by the Wright-Fisher model. It allows only two lineages at a time to combine, while in the discrete-generations Wright-Fisher model, more than two lineages can combine simultaneously since a single individual can have multiple offspring. Kingman derives his model by taking a series of discrete-generations Wright-Fisher models, with the kth of these having $N = k$ and a new time scale in which one unit of time is k generations. He shows that the limit of the genealogical processes of these models is one in which the (rescaled) time back to coalescence when there are n copies is distributed as

$$\tau \;\sim\; \mathrm{Exp}\left[4/(n(n-1))\right], \tag{1.2}$$

and he also shows that, in the limit, all coalescences are of only two copies. Returning to the original time scale, the limiting process approximates the genealogy specified by equation (1.1).

This sort of limit is well-known in theoretical population genetics – it is the one used to approximate gene frequency change by a diffusion process [12]. In effect, Kingman's n-coalescent is a diffusion approximation. Although diffusion processes approximate discrete changes of gene frequencies by a continuous diffusion process, they are extraordinarily accurate. One way that we can check this in the coalescent process case is to calculate whether coalescence will involve more than two lineages in the Wright-Fisher model. In the Wright-Fisher model, if we have n lineages and go back one generation, the probability that two copies coalesce while the others all do not will be $\binom{n}{2}$ times the probability that copies 1 and 2 coalesce and others do not, by the exchangeability of the process. As each copy chooses its ancestor independently, we need the probability that copy 2 chooses the same ancestor as copy 1, copy 3 chooses a different ancestor, copy 4 chooses an ancestor different from those two, copy 5 chooses an ancestor different from those three, and so on, so that the total probability of pairwise coalescence is

$$\binom{n}{2}\frac{1}{2N}\left(1-\frac{1}{2N}\right)\left(1-\frac{2}{2N}\right)\left(1-\frac{3}{2N}\right)\cdots\left(1-\frac{n-2}{2N}\right). \tag{1.3}$$

The probability that some of the copies coalesce is found by subtracting from 1 the probability that none coalesce, to get, by a straightforward argument:

$$1-\left(1-\frac{1}{2N}\right)\left(1-\frac{2}{2N}\right)\left(1-\frac{3}{2N}\right)\cdots\left(1-\frac{n-1}{2N}\right). \tag{1.4}$$

To first order, both of these expressions are equal, as both are

$$n(n-1)/(4N) \;+\; O(1/N^2) \tag{1.5}$$

which indicates that as N increases they become close, so that the probability that a coalescence involves more than two lineages becomes negligible. Taking the ratio of the expressions in equations (1.3) and (1.4), we can compute the fraction of coalescences that are coalescences of two lineages when there are 10 lineages for increasing values of N and get some sense of this (Fig. 1.1).

The fraction of two-way coalescences becomes high as the population size passes 100, which is the square of the number of lineages. We can also examine, for $N = 10,000$, the fraction of two-way coalescences with different numbers of lineages (Fig. 1.2).

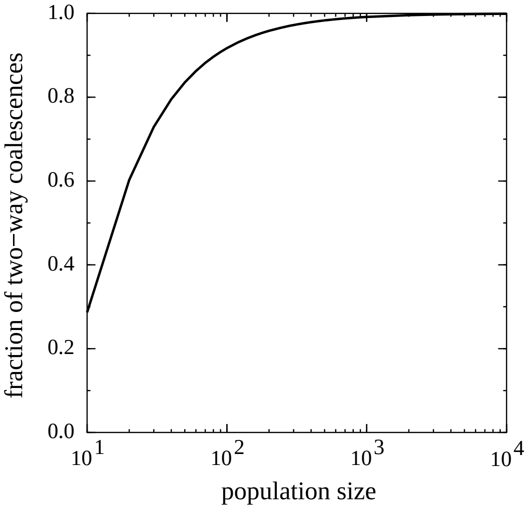

FIG. 1.1. The fraction of coalescences that are of two lineages, when there are 10 lineages, for different population sizes N.

These patterns can be summarized by saying that most coalescences will be two-way if $n^2 < N$. However it is not obvious that having a modest fraction of three- or four-way coalescences will invalidate inference methods that assume the coalescent, so the coalescent may be a good approximation even when this condition is violated.

The coalescent process predicts that the genealogy of copies in a population is a random branching tree. The coalescence times are individually exponentially distributed. The sum of their expectations is

$$\sum_{k=2}^{n} \frac{4N}{k(k-1)} \;=\; 4N \sum_{k=2}^{n} \left(\frac{1}{k-1} - \frac{1}{k} \right) \;=\; 4N \left(1 - \frac{1}{n} \right). \qquad (1.6)$$

We might expect that the total time for coalescence of the ancestors of a sample from a population is proportional to the sample size (or even to its square), but this calculation shows that it is actually almost independent of sample size.

One simple modification of this result is to use Sewall Wright's N_e in place of N. This quantity, the "effective population size" corrects for a variety of ways in which the mating system departs from a simple Wright-Fisher model. Formulas are available to calculate the appropriate corrections for separate sexes, unequal numbers of the two sexes, monogamy, overlapping generations, and variation of fertility from parent to parent. I will use N here, but the reader should keep in mind that N_e will usually be needed instead.

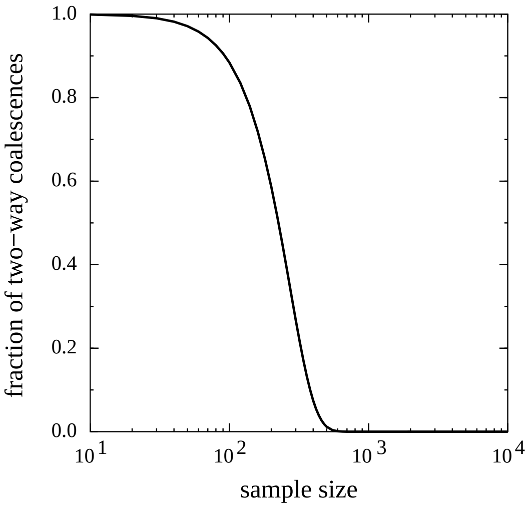

FIG. 1.2. The fraction of coalescences that are of two lineages, for different numbers of lineages, when population size $N = 10,000$.

1.2 Effects of evolutionary forces on coalescent trees

1.2.1 *Population growth*

The above theory is for a single population of constant size. When population sizes grow or shrink, the rate of coalescence changes. For example, if the population size is N for the most recent 500 generations, but before that is $N/10$ for 100 generations, and before that again N, the effect of this bottleneck on the coalescent is straightforward. Going back 500 generations, we have the usual coalescent process with rate (for k lineages) of $k(k-1)/(4N)$. If we get back to the most recent end of the bottleneck period and have at that time ℓ lineages, the rate of coalescence back beyond that is $10\,\ell(\ell-1)/(4N)$. If when the farthest end of the bottleneck is reached we have m lineages, the rate beyond that point is $m(m-1)/(4N)$. Thus there will tend to be a burst of coalescence at the time of population bottlenecks, though there may not be many coalescent events in those bottlenecks unless the length of the bottleneck in generations approaches the population size at that time. A bottleneck of population size of 1000 individuals may not have much effect if it lasts for only 10 generations.

It was noticed by Kingman [29] that there is a simple way to treat population growth if we can integrate the reciprocal of the population size. It makes use of the fact that a smaller population causes proportionately more coalescence per unit time. For example, if the population size N grows exponentially at rate g, the population size t generations ago was $N(t) = N(0)\exp(-gt)$. The rate of coalescence of k lineages t units of time ago would then be $k(k-1)/(4N(t)) = \exp(gt)\,k(k-1)/(4N(0))$. A coalescent process that has such time-dependent rates can be defined and simulated. A simpler way is to note that coalescence occurs $\exp(gt)$ times faster t units of time ago, because the population is that factor smaller then. It is as if the clock were running $\exp(gt)$ times as fast. We can change the time scale going backwards, to one that accumulates $\exp(gt)$ as much time t units of time ago. It has this fictional time be

$$\tau \;=\; \int_0^t e^{gu}\,du \;=\; \left(e^{gt}-1\right)/g. \tag{1.7}$$

On this fictional time scale, the coalescent process will have rates independent of time. The coalescent with an exponentially growing population is then simply the ordinary coalescent with population size $N(0)$, if we observe it on the fictional time scale τ. One can draw a random outcome of the coalescent process with exponential population growth by sampling the ordinary coalescent, considering the times of coalescence to be values of τ, and then computing the corresponding values of the actual time t by solving for t in equation (1.7) to get

$$t \;=\; \frac{1}{g}\;\ln(1+g\,\tau). \tag{1.8}$$

The effect of a positive growth rate g is to compress times in the past relative to the present. As Slatkin and Hudson [47] noted, the trees become closer to a "star tree" in which all lineages simultaneously radiate from a single node. If the growth rate is negative, the times at the base of the tree are stretched (sometimes infinitely so).

1.2.2 *Migration*

When we have more than one population, a coalescent tree forms in each population, but lineages also move between populations. Going backwards in time, if m_{ij} is the probability that a lineage in population i came from population j in the preceding generation, there is an event with probability $m_{ij}\,dt$ in the previous small interval of time of length dt. For example, if there were 3 populations of size N_1, N_2, and N_3, and if currently they contain respectively k_1, k_2, and k_3 lineages, the events that can occur during a small interval of length dt, going backwards in time, include coalescences within each of the three populations and migrations. The former happen with rates $k_1(k_1-1)/(4N_1)$, $k_2(k_2-1)/(4N_2)$, and $k_3(k_3-1)/(4N_3)$ per unit time. In population 1 there is a total rate $k_1m_{12}+k_1m_{13}$ of migrations, and similarly for the other two populations. The total rate of events for p populations is then

$$\sum_{i=1}^{p}\frac{k_i(k_i-1)}{4N_i} \;+\; \sum_{i=1}^{p}\sum_{\substack{j\,=\,1\\ j\,\neq\,i}}^{p} k_i m_{ij}. \tag{1.9}$$

To draw a genealogy from the coalescent with migration, we proceed backwards in intervals. We draw the length of the interval from an exponential distribution whose mean is the reciprocal of the quantity in 1.9. We then decide

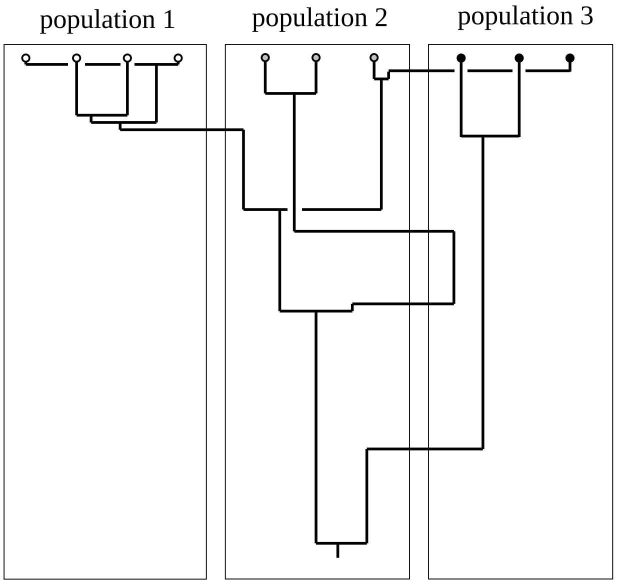

FIG. 1.3. A simulated coalescent with migration among adjacent populations with three populations of equal sizes and $4Nm = 1$ in each, going backwards from samples of 4, 3, and 3 lineages.

whether the event is a coalescence or a migration, by drawing these in proportion to their total rates of occurrence, and then we decide in which population each event is and which lineage or lineages it involves.

Figure 1.3 shows a randomly sampled coalescent from three populations of equal size N, who have symmetric migration with $4Nm_{ij} = 1$.

The coalescent process for migration was first investigated by Takahata [50] and (somewhat implicitly) by Hudson and Kaplan [27] and by Kaplan, Darden, and Hudson [28].

1.2.3 *Coalescents with recombination*

So far we have assumed that each copy of a gene is descended from a single copy in the preceding generation. This is true if there is no genetic recombination within the gene. If there is recombination possible, the copy could be descended from both copies in the parent. At any one site in the DNA sequence, the gene is descended from only one copy, and the coalescent at that site is the normal one. But when the sites are taken together, the genealogy is not a tree. When we approximate the genealogy of the sequence by a coalescent, recall that in effect we consider cases with large population size N, and small rates of such forces as migration. To obtain a coalescent approximation to a recombining genealogy, we also take the recombination rate per site per generation, r, to be small. This means that we will assume that there cannot be more than one recombination event in a sequence in one generation.

To model recombination, we assume that when a recombination event occurs in a sequence which has L sites, it does so at one of the $L-1$ intervals between sites, chosen at random. The sequence before the point of recombination comes from one of the two parental copies, and the sequence after the point of recombination comes from the other parental copy. The two copies that are in the parent are themselves drawn at random from the population, so they go back in time along independent lineages that can coalesce with others, or even with each other. In tracking the ancestry of a population sample, we will want to have each lineage accompanied by a set S of sites. In the sample, the sets S are all $\{1, 2, \ldots, L\}$. As the lineages go back in time, they have the usual probabilities of coalescing and migrating. There are also recombination events occurring stochastically at rate $4Nr$ per interval between adjacent sites. When a recombination event occurs, if it occurs just after site ℓ it divides the set of sites into two subsets, $\{1, \ldots, \ell\}$ and $\{\ell+1, \ldots, L\}$. The set of sites "active" in the two parent haplotypes are then changed to $S \cap \{1, \ldots, \ell\}$ and $S \cap \{\ell+1, \ldots, L\}$. When two lineages coalesce, the set of active sites is the union of the two sets of active sites, though the set of intervals available for recombination is from the leftmost site in that union to the rightmost site.

We can represent the genealogy by a graph called the *ancestral recombination graph* [24,20]. Figure 1.4 shows an ancestral recombination graph with three tips, four coalescences (the shaded circles) and two recombination events (the white circles). Next to each line is the list of sites in that lineage (out of a total sequence length of 1000) that are "active" in the sense of being ancestral to sites in the tip sequences. Note that one lineage has a disjoint list of active sites.

An alternative way of thinking of genealogies with recombination is to think of the genealogies at the different sites. At each site the genealogy is a simple coalescent. Neighboring sites between which there has been no recombination have the same coalescent. In the example in Fig. 1.4 the first 265 sites have one coalescent tree, the next 127 sites another, and the final 608 sites a third. Wiuf and Hein [56] have defined a stochastic process that makes changes in the coalescent as one moves along a sequence in a way that correctly generates an ancestral recombination graph. Most computer simulation of ancestral recombination graphs uses the program of Hudson [26] which generates the graph by moving backward in time and considering the sets of sites in different lineages.

It is helpful to have a sense of the rate at which the coalescent tree changes as one moves along the genome. How far must we go to have the tree be effectively independent? A simple calculation can be based on the distance we must move along the genome so that a lineage from a tip down to the root of the coalescent tree is expected to have one recombination event. The distance to the root is close to $4N$ generations. So we want to find how far along the genome we must go to have $4Nr = 1$. In a human meiosis there is about one recombination event per 10^8 bases. If the effective population size tens of thousands of years ago were

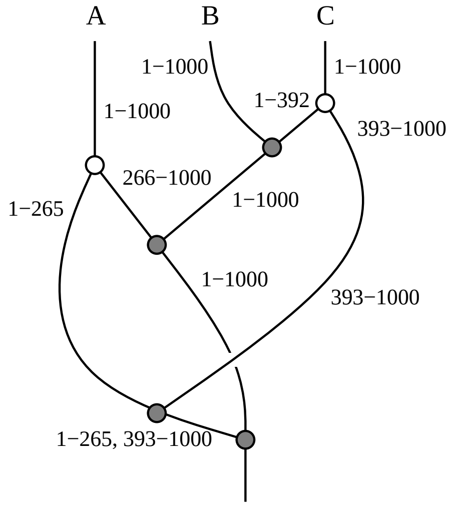

Fig. 1.4. An ancestral recombination graph for a sample of three sequences of 1000 bases. Next to each lineage are listed the sites in it that are ancestral to the tip sequences. Coalescent events are shown as shaded circles, recombination events as white circles.

10^4, and the recombination rate were the same throughout the genome, this implies a short distance, 2500 bases. If the effective population size were higher, say 10^5, the distance is even shorter, only 250 bases!

You may wonder what justification I have for the rule $4Nr = 1$. In fact, the condition for similarity of trees is the same as the condition for there to be nonrandom association of alleles at loci. These associations are known as linkage disequilibrium. The coalescent tree at one site strongly affects the distribution of alleles in the sample. An allele that has arisen by mutation at that site tends to occur in the descendants of a single branch of the coalescent tree. If another site shares the same coalescent tree, one of its alleles will be strongly positively or negatively associated with the allele at the first site. Robertson and Hill [45] make a calculation closely similar to the above one, calculating the size of blocks of linkage disequilibrium.

Models can also be made of the effect of gene conversion on the coalescent, although as yet there has been little use of them.

1.2.4 *Natural selection*

It has been difficult to accomodate natural selection in coalescents, but recently there has been some progress in doing so. If there is no natural selection occurring, then the shape of the coalescent genealogy is not affected by which copies have which DNA sequence. In the presence of natural selection, there is such a dependence. If we have (say) 5 copies of one allele, and 5 of another, and if the first allele has higher fitness than the second, then most likely the first allele is spreading in the population. If so, it is more probable that two copies of it coalesce when we go back in time than two copies of the other allele. The result is that we cannot specify any coalescent without knowing more about the DNA sequence in the copies.

For many years this was thought to make it impossible to specify any coalescent process in the presence of natural selection. Krone and Neuhauser [40, 31] discovered a way to do so. It creates a coalescent by going back in time and having both coalescence events and also special forks that reflect a natural selection event. This produces a genealogy with loops in it. The genotype is then specified at the root of this genealogy, drawn from an appropriate population-genetic equilibrium distribution. Then genotypes are propagated up the genealogy, allowing for mutation events as well. When the top of a loop is reached, it is decided which side of that loop connects upward, depending on its genotype. Krone and Neuhauser's result is a breakthrough, though it does not specify a genealogy independent of the genotypes of the gene copies, as the other coalescent processes do.

Earlier treatments of natural selection [27, 28] could handle only cases of strong natural selection, which in effect divides the copies into subpopulations whose sizes are the consequence of the fitnesses.

1.3 Inference methods

Having understood the stochastic processes that produce treelike genealogies of gene copies, the next obvious step should have been to find a way to use these to compute likelihoods or carry out Bayesian inference of parameters. The central model framework for doing so is the neutral mutation theory of genetic variation, widely studied since the 1960s. Molecular sequences have been modelled as evolving under genetic drift and mutation, without natural selection. This model also serves as a null hypothesis against which to test for the presence of natural selection.

In a coalescent, mutation can be accomodated by allowing it to occur on the branches, modelled as happening in continuous time. This is the same model used in the inference of phylogenies. The difference is that in the coalescent case, the coalescent genealogy is not being estimated, but instead is part of the machinery of statistical inference of the population and genetic parameters.

The models of mutation used are the usual models of sequence mutation used with phylogenies. The presumption in most cases is that the mutations are selectively neutral, with no fitness differences. Two approximate models are also in wide use in the population genetics literature. One is the *infinite alleles model*, due to James F. Crow and Motoo Kimura [4]. In it there is a constant risk of mutation, at rate μ per locus, to a completely new allele. All alleles can be distinguished, but they give us no clue which ones are derived from which other ones. The same allele never arises twice. Mutations in DNA sequences behave approximately like this, as long as there are so many sites that the chance of the same site mutating again is small. However, in real DNA sequences, the sequence does give us information about which sequences are likely to be separated by one mutational event.

A closer approximation is the *infinite sites model* of Watterson [52]. It represents the gene by a line segment, and each mutation occurs at a random location chosen from (0,1). As such, no mutation ever recurs at the same exact location. It is assumed that we can see the line and the placement of the differences, but it is also usually assumed that we cannot know, at a site which has a variation, whether the presence or absence of the variation is the original state. Thus, if we see three copies that have their lists of variations present as $\{0.366, 0.8197\}, \{0.366\}$, and $\{0.684\}$, the variation counted as present at position 0.366 in the first two copies could also be considered as one that is absent in those copies but present in the third. The lists would then be $\{0.8197\}$, $\{\}$, and $\{0.366, 0.684\}$. If the variation at position 0.684 was considered absent in the third copy but present in the other two, the lists would be $\{0.684, 0.8197\}, \{0.684\}$, and $\{0.366\}$. These are all completely equivalent. As long as there is no recombination allowed within the locus, the exact locations on the line segment actually do not matter, and each mutational event in effect partitions the copies into two sets. The partitions are ordered and are compatible, in that when we intersect any two such partitions they form no more than three sets. We shall see the infinite sites model used in some of the inference methods below.

1.3.1 *Earlier inference methods*

It is a puzzling fact that little attention was paid to likelihood inference (and Bayesian inference) in population genetics until the 1990s. Some of this inattention may have been the result of the apparent intractibility of the problem. The only model for which a likelihood could be computed was Ewens's [9] model of a locus undergoing mutation and genetic drift under an infinite-alleles model of mutation. (One should mention also R. C. Griffiths for deriving a likelihood inference of population divergence time under that same model [18]). But one would have thought that the problem would at least have been posed as a major challenge for theoretical population geneticists. It was not.

This may be related to the high prestige in that field of closed-form solutions for distributions and changes of population composition, and the correspondingly low prestige of statistical and computational methods. For example, for a field with so much mathematically sophisticated theory, population geneticists maintain relatively few web sites and distribute relatively few computer programs. They are far outclassed in this by systematists and molecular evolutionists, even though those fields are mathematically less sophisticated. Although likelihood and Bayesian inference methods became dominant in statistical inference from human pedigrees during this period, population geneticists working on evolution tended to ignore the likelihood paradigm and instead derive expectations and variances for particular statistics.

Many of those were heterozygosities which involved first and second moments of gene frequencies. These can be shown to lose statistical power compared to coalescent-based methods [13,16]. Another widely-used statistic for the infinite-sites model, Watterson's number of segregating sites [52], is more powerful, but still less so than likelihood-based methods [13,14,16].

1.3.2 *The basic equation*

The first key to computation of the likelihood for a population sample of molecular sequences is that we can compute it straightforwardly once the coalescent tree is known. The likelihood models of phylogenetic inference allow the computation of $\mathrm{Prob}\,(\mathrm{D} \,|\, \mathrm{T}, \mathcal{P})$, the probability of the sequences given the tree and the values of the relevant parameters. The second key is the realization that we do not know the tree T, but that the sequences do give us some information about it. The likelihood $\mathrm{Prob}\,(\mathrm{D} \,|\, \mathcal{P})$ is

$$\begin{aligned} \mathrm{Prob}\,(\mathrm{D} \,|\, \mathcal{P}) &= \sum_{T} \mathrm{Prob}\,(\mathrm{D}, \mathrm{T} \,|\, \mathcal{P}), && (1.10) \\ &= \sum_{T} \mathrm{Prob}\,(\mathrm{D} \,|\, \mathrm{T}, \mathcal{P})\ \mathrm{Prob}\,(\mathrm{T} \,|\, \mathcal{P}). && (1.11) \end{aligned}$$

The summation is over all possible coalescent trees, and includes not only summation over tree topologies but integration over all possible combinations of coalescence times. The first term inside the summation is easily computed by the standard dynamic programming methods of phylogeny inference. The second is the density of the coalescent distribution.

1.3.3 *Rescaling times*

In the simplest case, of one population, the parameters in equation (1.11) are the population size, N, and the mutation rate per site, μ. In fact, they cannot be inferred separately. If we change the time scale of the branch lengths of the tree T so that they are given, not in generations, but in units of expected mutations per site, the expression for the likelihood now becomes a function of the product $4N\mu$ and the quantities μ and N do not appear separately. This makes intuitive sense – if we are computing the joint probability of a set of sequences observed at the present, there will be no difference between a tree with a given mutation rate μ and one which is twice as deep but has half the mutation rate. The depth of the tree is proportional to N, so that the likelihood is a function only of the product $N\mu$. It is a convenience to express the product as $\Theta = 4N\mu$.

In this simple case, the likelihood can then be written as

$$\mathrm{Prob}\,(\mathrm{D} \,|\, \Theta) = \sum_{\mathrm{G}} \mathrm{Prob}\,(\mathrm{G} \,|\, \Theta)\ \mathrm{Prob}\,(\mathrm{D} \,|\, \mathrm{G}) \qquad (1.12)$$

since the branch lengths of the coalescent genealogy G are now expressed in mutational units.

The sum is of a product of two terms. The first is the coalescent density. If the ith coalescent interval on the tree G is u_i, measured in mutational units, then the coalescent density for n sequences is

$$f(G \,|\, \Theta) \;=\; \prod_{i=1}^{n-1} \left(e^{-\frac{(n-i+1)(n-i)}{\Theta}\, u_i} \;\frac{2}{\Theta} \right). \tag{1.13}$$

The density is easy to calculate once we know the u_i. Likewise the second term on the right-hand side of equation (1.12) is easy to compute, using the standard recursion for likelihoods on phylogenies. Although likelihood methods can be slow, this is not so much true for the computation of the likelihood for one tree, as we have one topology and are not optimizing the branch lengths.

1.3.4 *How many coalescent trees?*

This would seem to solve the problem, except for one matter. The summation is over all possible coalescent trees that could connect the sequences. Each tree is specified by a given sequence of pairs of lineages that coalesce, plus the times of these coalescences. With n lineages, the sequence of coalescence events is specified by choice of pairs of lineages to coalesce. The total number of possibilities is

$$\prod_{i=1}^{n-1} \binom{n-i+1}{2} \;=\; \frac{n!\;(n-1)!}{2^{n-1}}. \tag{1.14}$$

These different possibilities are called labelled histories – they are different trees in which we distinguish between the order of interior nodes in time. They were defined by Edwards [8]; the formula counting them is given in that paper.

The number of labelled histories rises rapidly, more rapidly than the number of tree topologies. For only 10 tip species, there are 2,571,912,000 of them. Worse yet, evaluating the likelihood involves integrating over all possible coalescence times. There are $n-1$ of these, so for 10 tips we must evaluate 2.571×10^9 integrals, each 9-dimensional. It would be a great economy if there were a closed-form formula for the integration, but there has been no progress toward that.

1.3.5 *Monte Carlo integration*

The integral in equation (1.12) can be thought of as the expectation of $\mathrm{Prob}\,(\mathrm{D} \,|\, \mathrm{G})$ over the Kingman coalescent distribution for parameter value Θ. If we cannot do the integrals analytically, and cannot hope to do them all numerically, a natural alternative is Monte Carlo integration. Perhaps we can draw a large sample of coalescent genealogies from the Kingman density, compute $\mathrm{Prob}\,(\mathrm{D} \,|\, \mathrm{G})$ for each, and average.

I have tried to implement this at least once, and the results were disastrous. For almost all of the possible genealogies G the value of $\mathrm{Prob}\,(\mathrm{D} \,|\, \mathrm{G})$ is nearly zero; for a small minority it is much larger. The result is that the averages vary wildly from one sampling run to another, and no accurate estimate of the overall likelihood is obtained.

1.3.6 *Importance sampling*

It thus becomes essential to find some way of concentrating the sampling in the relevant regions. The correction that needs to be made for importance sampling has long been known. If we want to compute the expectation of function $h(x)$ over a distribution whose density function is $f(x)$, but we choose the samples from a distribution whose density function is $g(x)$, it is easy to see that

$$\begin{aligned} \mathbb{E}_f\left[h(x)\right] &= \int_x f(x)h(x)\,dx, &(1.15)\\ &= \int_x \frac{f(x)}{g(x)}\,g(x)\,h(x)\,dx, &(1.16)\\ &= \mathbb{E}_g\left[\frac{f(x)}{g(x)}\,h(x)\right]. &(1.17) \end{aligned}$$

We correct for the importance sampling by averaging, not $h(x)$ but $(f(x)/g(x))h(x)$. An intelligent choice of the density $g(x)$ can concentrate our sampling on coalescent trees that make a substantial contribution to the integral. The factor $f(x)/g(x)$ corrects for the excessive density of points in some areas of the space. If, for example, $g(x)$ concentrates twice as many sampling points around x as $f(x)$ would, the factor $f(x)/g(x)$ weights the samples to reflect the fact that each should be taken to represent half as much area in the space as it would if we sampled from the density $f(x)$.

Importance sampling makes numerical sampling approaches to likelihood inference or Bayesian inference with coalescents practical. Methods have been developed that draw independent samples, and also methods that draw correlated samples. I will call both of these “sampling methods”. With the rise in popularity of Markov chain Monte Carlo (MCMC) methods as means of sampling from difficult distributions, it was inevitable that they would be applied to this task. Although the drawing of independent samples is a trivial case of a Markov chain, designation as MCMC methods is usually reserved for the correlated samplers.

1.3.7 *Independent sampling*

The pioneers in applying sampling methods for computing likelihood functions in coalescents were Griffiths and Tavaré [21]. For samples whose mutational process was the infinite sites model, Griffiths [19] had envisaged using a recursion (due to Golding [17]) to compute all possible sequences of mutational and coalescent events that could have led to the observed sample. This proved to be too difficult computationally for more than a few samples. Griffiths and Tavaré [21] proposed instead sampling paths through the recursion, and for each computing a functional that reflected the probabilities of events. Each such path is an independent sample, a very desirable property, as it thus completely avoids the problem of getting stuck in one region of the space.

At each stage, Griffiths and Tavaré consider the possible events that could happen (going backwards in time). If there is only one sequence that has a particular site in the mutant state, then it is possible that this event is a mutation. If there is more than one copy of a sequence, it is possible that this event is a coalescence of two of them. They sample these events proportional to their probability of occurrence, but not allowing those that would conflict with the data. Suppose that there was one sequence that carries a mutant allele at position 0.2, another with mutant alleles at positions 0.4 and 0.5, and a third with a mutant allele at position 0.2. With three sequences, we could have three possible coalescences, and there are four copies of the mutant that could have recently mutated (so that going backwards they unmutate). But as we have an infinite sites model, position 0.2 cannot unmutate in either of its positions (i.e., the most recent event cannot have been a mutation creating that mutant allele). Of the three possible coalescences, two of them could not have been the most recent event, as the genotypes of those pairs of sequences are different. In such a case, Griffiths and Tavaré sample from among the one allowable coalescence and two allowable mutations in proportion to their probabilities.

Griffiths and Tavaré go back in time, sampling possible events, until the sample coalesces to one sequence. They then compute a functional, which is simply the appropriate importance sampling weight. Their method can either be thought of as sampling paths through the recursion, or sampling sequences of past historical events. These are equivalent. The events define a genealogical tree with mutations indicated on it, but no time scale is needed.

There is one more subtlety. We can't actually know for any site that shows variation in our sample which of its two states is the original state and which the mutant. So Griffiths and Tavaré, in computing their importance sampling weights, use the probabilities of unrooted trees rather than of rooted trees, in effect summing up over all the ways that the ancestral state at the individual sites could be interpreted.

I have given a rather cursory description of their method here – a more detailed consideration of the way it fits into the framework of importance sampling is given by Felsenstein *et al.* [15].

This independent sampling (IS) method is attractive because it not only entirely avoids getting stuck in regions of tree space, but each sample is rapid. However, because the importance sampling is imprecise, it often needs large numbers of samples to be sure of sampling from the trees that contribute most of the probability. It also approximates the mutation process by an infinite sites model, which means that sites at which there are back mutations or parallel mutations must be removed from the data to avoid getting a likelihood of zero.

The original sampler allowed for either constant or exponentially growing populations. Bahlo and Griffiths [1] have extended the method to multiple populations with migration, and Griffiths and Marjoram [20] have extended it to sampling of ancestral recombination graphs.

The IS sampler can be extended to models of DNA sequences, but it then proves extremely slow owing to the high probability that mutations going backwards in time will lead to widely divergent sequences. This problem was addressed by Stephens and Donnelly [48], who have speeded up the IS sampler by a large factor in the DNA case by biasing the sampling of mutations in different sequences toward tracing back to a common ancestral sequence, and making the appropriate importance sampling correction. De Iorio and Griffiths [5] have derived an independent sampling method from consideration of the diffusion approximation. They show that this leads directly to Stephens and Donnelly's method, which thus can be seen to be a particular case of a more general approach. They also [6] extend their method to subdivided populations with migration among them. This approach can presumably be used as a general method for developing efficient independent sampling methods for other mixtures of evolutionary forces.

Fearnhead and Donnelly [10] have made another such correction that greatly speeds up independent sampling in the case of recombination, making it much more practical. They have presented simulation evidence that their independent sampler performs better than the correlated sampler described below.

1.3.8 *Correlated sampling*

A second approach by Kuhner *et al.* [34] comes from our lab. We sample our way through tree space by sampling coalescent genealogies. In the simple case of estimating Θ in a population of constant size, we used a trial value, the "driving value" Θ_0, and wanted to achieve an importance sampling distribution whose density function was proportional to $\mathrm{Prob}\,(\mathrm{G} \,|\, \Theta_0)\ \mathrm{Prob}\,(\mathrm{D} \,|\, \mathrm{G})$. If Θ is close to Θ_0, this would be nearly an optimal choice. Using equations (1.12) and (1.17), if we are trying to compute the likelihood, it will be the average over sampled trees of

$$\mathrm{Prob}\,(\mathrm{G} \,|\, \Theta)\mathrm{Prob}\,(\mathrm{D} \,|\, \mathrm{G}) \Bigg/ \left(\frac{\mathrm{Prob}\,(\mathrm{G} \,|\, \Theta_0)\mathrm{Prob}\,(\mathrm{D} \,|\, \mathrm{G})}{\int_{\mathrm{G}} \mathrm{Prob}\,(\mathrm{G} \,|\, \Theta_0)\ \mathrm{Prob}\,(\mathrm{D} \,|\, \mathrm{G})} \right). \tag{1.18}$$

The denominator of the denominator is simply the likelihood at Θ_0, so after some cancellation this is

$$\frac{\mathrm{Prob}\,(\mathrm{G} \,|\, \Theta)}{\mathrm{Prob}\,(\mathrm{G} \,|\, \Theta_0)/\mathrm{L}(\Theta_0)}. \tag{1.19}$$

If we sample n genealogies $G_1, G_2, \ldots G_n$ in our Markov chain Monte Carlo run, and average this quantity, we find that $L(\Theta_0)$ can be factored out so that

$$\frac{L(\Theta)}{L(\Theta_0)} \;=\; \frac{1}{n}\sum_{i=1}^{n} \frac{\mathrm{Prob}\,(\mathrm{G_i} \,|\, \Theta)}{\mathrm{Prob}\,(\mathrm{G_i} \,|\, \Theta_0)}. \tag{1.20}$$

Thus the likelihood ratio between Θ and Θ_0 is estimated by the mean ratio of the Kingman coalescent densities for each tree at these two parameter values. The reader may wonder what happened to the data, which appears nowhere in equation (1.20). Its influence is felt entirely through the sampler that chooses the G_i.

1.3.8.1 *Tree proposals* To implement this sampler, we need a proposal mechanism and the usual Metropolis/Hastings acceptance-rejection method. Although we initially used a much more limited tree rearrangement method, the proposal mechanism we have found most useful (invented by Peter Beerli) is to choose a node in the coalescent tree (excluding the root), and then dissolve the connection between it and the node immediately ancestral to it. This lineage is then allowed to reconnect to the tree by a conditional coalescent. A conditional coalescent is a distribution whose density is proportional to the coalescent in all regions where it is not zero. We sample from this by having the lineage go back in time, having at any moment when there are k other lineages an instantaneous rate k/Θ_0 of coalescing with a random one of them. The lineage finally hooks itself back into the tree. This can result either in a small change of the time of the coalescent node or a major relocation of the lineage in the tree.

The Metropolis-Hastings sampler for this conditional coalescent proposal mechanism turns out to be to accept the new genealogy with probability

$$\min\left[1,\ \frac{\mathrm{Prob}\,(\mathrm{D}\,|\,\mathrm{G}_{\mathrm{new}})}{\mathrm{Prob}\,(\mathrm{D}\,|\,\mathrm{G}_{\mathrm{old}})}\right]. \tag{1.21}$$

The terms for the Kingman coalescent are cancelled by the Hastings correction for the biased proposal mechanism. This is convenient but not a large computational saving. The computations in 1.21 are still considerable, much more than for sampling a single event history in the independent sampler.

The sampler does considerably better if Θ_0 is close to the true Θ. In our programs, we run an MCMC chain, infer a new value of Θ, and use that as Θ_0 for the next chain. In a typical run, we do this 10 times, then use the resulting Θ as the basis for one longer chain to get an even more accurate Θ. This in turn is used for one final long chain to infer the likelihood ratio curve and the final estimate of Θ.

1.3.8.2 *Advantages and disadvantages* The correlated sampler has some obvious disadvantages. It could become stuck in one region of the tree space, and the calculations for each sample are much larger than for the independent sampler. However, there are advantages as well. If Θ_0 is close enough to Θ, the trees sampled are close to being an optimum sample of the trees proportional to their contribution to the likelihood. The independent sampler is less accurate, and that can lead it to need much larger numbers of samples than the correlated sampler. No clear conclusion has emerged about which method is superior.

1.3.8.3 *Extensions of the correlated sampler* Like the independent sampler, the correlated sampler has been applied to more complex cases. Kuhner *et al.* [35] have incorporated exponential population growth, Beerli and Felsenstein [2, 3] have incorporated migration among a number of populations, and Kuhner *et al.* [36] have incorporated recombination by having the sampler move in a space of ancestral recombination graphs.

One interesting discovery was made in the course of the work on exponential growth. It had been overlooked in previous coalescent studies. It was found [35] that the estimate of growth rate is strongly biased toward positive growth. If we estimate both Θ and the scaled growth rate g/μ, the maximum likelihood estimate of growth rate would usually be strongly positive even when true growth rate was 0. This behavior is less alarming when it is considered that the interval of allowable growth rates is wide in these cases, and quite frequently contains 0 as well. The reality of this bias can be demonstrated in the case of a sample size of two sequences, when the integration can be done numerically without MCMC sampling. The bias is little reduced by adding more samples, but is strongly reduced by adding more loci. That allows us to rule out the possibility of a strong positive growth rate by occasionally finding loci with deep coalescences.

1.3.9 *Sampling from approximate distributions*

The computational difficulty of the sampling methods has led to the development of approximate methods that try to retain much of the statistical power of the exact samplers, while avoiding all or most of the sampling effort. This has been particularly tempting in the case of recombining coalescents, where the size and complexity of the ancestral recombination graph is daunting. Li and Stephens [37] have introduced the PAC (product of approximate conditionals) likelihood method for inferring the recombination from a sample of haplotypes. This approximates the coalescent distribution for the sample as the product of conditional distributions, each itself an approximation. The resulting calculation is far faster than any of the sampling approaches. It has become widely used. Hudson [25] and McVean *et al.* [39] have both used a different approximate method, one which approximates the distribution of haplotypes as the product of two-locus distributions. Fearnhead and Donnelly [11] give another approximate method based on using sampling methods on subregions and deriving an approximate likelihood from the results. Li and Stephens present simulations comparing these methods, finding that their method does best.

Those methods make an approximate computation of the likelihood of the full data. An alternative approach is to reduce the data to some appropriate summary statistics, and compute the likelihood for those reduced data. This was pioneered by Weiss and von Haeseler [53]. A more extensive consideration of methods for approximate inference that do not involve computing the full likelihood of the full data is given by Marjoram *et al.* [38]. While these methods enable much more rapid computation, the issue that must always be kept in mind is whether the summary statistics retain enough information.

1.3.10 *Ascertainment and SNPs*

The growth in the use of SNP (single nucleotide polymorphism) data has raised another issue, ascertainment bias. If sites are screened and only those found to be varying in some panel of genomes are included, we will find these sites to be much more variable in our sample than randomly sampled sites would be.

If we included these sites without making any correction for the screening, the result would be an unrealistically high estimate of the mutation rate μ. That in turn would lead us to misestimate the rates of other parameters – for example, discrepancies in the picture of the tree from different sites that might actually be a sign of recombination would instead be too readily attributed to recurrent mutation.

Several papers have derived the corrections needed for the ascertainment of SNPs [42, 32, 6]. Both treat various possible ways in which a SNP screening panel could be chosen. However, neither is able to treat the horrible reality. In some cases, ethical or legal concerns prevent the release of enough information about the panels to enable any sensible ascertainment correction to be made. The data are thus safe from being abused, and also safe from being used. Until recently, large-scale genomics projects acted as if they were blissfully unaware that analysis of their data required knowledge of how the screening was done. They either did not release the required information or, in some cases, they simply did not know it, or know that they had to know it. For some purposes (such as using the SNPs for linkage studies in pedigrees) this may not matter, but for all population analyses it matters a great deal. It is gradually beginning to be realized that an inability to correct the data for the way in which sites were chosen rules out many important uses of the data, making them largely a waste of money.

1.3.11 *Bayesian samplers*

I have so far discussed only likelihood inference. The spread in the popularity of Bayesian inference has led it to be applied to coalescent-based inferences [54, 55, 7]. In Bayesian sampling one updates both the genealogy and the values of the parameters, sampling from these in proportion to their contribution to the posterior distribution of the parameter values. This can involve simultaneous updates of parameters and trees, or it can involve alternating updates of parameters and trees. The technology of sampling is very similar to the correlated sampler, but the use of the resulting sample is very different. In the likelihood-based methods, one uses the samples of the trees to compute a likelihood curve. In Bayesian methods one uses the sample of parameter values as a sample from the desired posterior, while ignoring the trees.

Bayesian samplers are attractive in their simplicity. They also have a tendency to avoid problems with driving values, as they sample broadly from the possible values of the parameters. When the objective is not Bayesian, these samplers can still be usefully employed and the posterior distribution of parameters ignored. One issue with posterior densities of parameters is that we need some means of interpolating density between the sampled parameter values. This leads to convolution of the extremely spiky posterior distribution with broader kernels that smooth out the density. All these are to some extent arbitrary.

As with likelihood methods, approximate calculations and use of summary statistics rather than the full data enable much faster computation. The Approximate Bayesian Computation (ABC) method of Tavaré and his coworkers [44,38] takes advantage of this with, as is inevitable, the concomitant worries about whether one has chosen the best summary statistics.

1.3.12 *Future extensions*

I have barely skimmed the surface of the very active literature on coalescent-based inference. Coalescent methods are continually expanding. They will ultimately deal with all issues in evolutionary genetics. Some of the major extensions of the methods under way are:

Sequential sampling Coalescent methods have assumed that all samples are contemporary. If we can sample DNA from the past, some samples are at different levels in time in the tree. These need to be scaled using the mutation rate per generation (μ) and the generation time (T) to put them on the scale of branch length. In the simplest case [46], of the three quantities N, T, and μ, we can estimate two of them. This is an improvement over the case of contemporary tips, where we can only estimate one of these quantities, the product of N and μ. Sequential sampling is important in studies of ancient DNA, and is even more widely used in studies of rapidly evolving viruses such as HIV, where samples from the same patient over time must be considered to be at different levels of a tree. Sequential sampling methods are starting to be available in widely-distributed programs [7]. For a more extensive treatment of sequential sampling coalescent methods see the chapter by Rodrigo, Ewing and Drummond in this book.

Uncertainty about haplotypes Data frequently come as diploid genotypes. The usual way of handling these has been to try to resolve haplotypes, then treat those reconstructed haplotypes as if they were observations. A more realistic treatment would be to sum the likelihoods for all possible haplotype resolutions, so that we incorporate our uncertainty about the haplotype resolution into our statistical analysis. This has been proposed by Kuhner and Felsenstein [33]. It requires extra rounds of MCMC sampling, as we sample from among all possible haplotype resolutions. The method is not available in most distributed programs – when it is, it may replace most haplotype resolution calculations.

Multiple species It has been known since the work of Tajima [49] and Takahata and Nei [51] how to extend the coalescent to multiple related species. Each lineage in a tree of species will have a coalescent inside it, and such coalescents at different loci are independent of each other. If we arrive at a common ancestor, any gene copy lineages in each species that are not yet coalesced (going backwards in time) now join a common pool and are available to coalesce with each other. (It is best not to think of these matters forward in time, and thus not to use the confusing concept of "lineage sorting"). Likelihood and Bayesian treatments of inferences about species trees from single and multiple loci have begun to appear [41,43] and to be made available in computer programs [7,55]

Linkage disequilibrium mapping . It is customary in genomics for researchers to debate which measure of linkage disequilibrium to use to characterize the joint distribution of variation at linked sites. The correct answer is "none of them". As we have seen, trees and D's are intimately related, and multiple-locus linkage disequilibrium describes the same phenomena as do trees of recombining haplotypes. While the two equivalent descriptions can be interconverted, it is the coalescent description that is easier to work with. For a fully powerful analysis of multiple linked sites, the correct way to compute the location score is to compute the likelihood for each possible location of the disease locus. A Bayesian approach might propose different locations for the disease locus, but it would accept or reject these based on these likelihoods. In either case one needs a full coalescent calculation. This point has been realized by all major researchers on recombining coalescents, but it has taken some time for linkage disequilibrium mapping methods based on coalescents to become available. That situation is about to change, and the discussion of methods in genomics will change with it.

Selection Inferring locations in the genome where there may have been selective sweeps or where there may be balanced polymorphisms is possible by likelihood or Bayesian methods. To do so, natural selection needs to be incorporated into the coalescent framework. This is perhaps the most interesting frontier of coalescent methods; it is under active exploration by a number of groups. As coalescent methods for detecting selection become widely available, they should replace the present summary-statistics methods.

Inferring the history As we sample past coalescent histories of our data, we can see historical events such as the times of particular coalescences. We could also imagine reconstructing when particular mutations occurred [22]. Knowing exactly what happened in the past has great appeal, and is always of interest to the popular science media. Taking a reasonable sample will usually show these inferences to be very noisy. In addition, they are not inferences of the parameters of the underlying models. As such, they are not maximum likelihood estimates, but rather maxmimum posterior probability estimates (in a Bayesian framework they have posterior probabilities just as do the parameters). The question arises: is reconstructing the exact history a trivial pursuit? The quantities which are needed in further analyses are usually the underlying parameter values rather than the exact times of particular events. However, the ages of mutations or the depths of particular coalescences can serve as indications of whether an allele is not neutral, or a population size not constant. The jury is not yet in on how interested we should be in these reconstructions of history.

1.4 Programs

There are now many coalescent programs available. As of the summer of 2006, some of the main ones I am aware of are:

LAMARC Likelihood-based inference including inference of migration, population growth, and recombination.
`http://evolution.gs.washington.edu/lamarc.html`

GENETREE Maximum likelihood estimation of mutation, migration and population growth parameters and inference of times of coalescence and of mutation.
`http://www.stats.ox.ac.uk/∼griff/software.html`

BEAST Bayesian estimation of population sizes and growth rates, allowing for sequential sampling. Allows a “relaxed” molecular clock.
`http://evolve.zoo.ox.ac.uk/beast/`

BATWING (Bayesian Analysis of Trees With Internal Node Generation) Bayesian inference of mutation and population growth, with single or subdivided populations.
`http://www.mas.ncl.ac.uk/∼nijw/`

msvar Bayesian inference of mutation rate and growth rate from microsatellite data for multiple loci in one population.
`http://www.rubic.rdg.ac.uk/∼mab/software.html`

MDIV Likelihood inference of divergence time and migration rates for two populations.
`http://www.binf.ku.dk/∼rasmus/webpage/mdiv.html`

MICSAT Likelihood inference for single-step microsatellite models.
`http://www.mas.ncl.ac.uk/∼nijw/#micsat`

MISAT Likelihood inference of mutation rates for single- and multi-step models of microsatellite evolution in a single population.
`http://www.binf.ku.dk/∼rasmus/webpage/misat.html`

IM (Isolation with Migration) Likelihood inference of divergence times and effective population sizes in a model with two diverged populations with subsequent migration between them.
`http://lifesci.rutgers.edu/∼heylab/HeylabSoftware.htm#IM`

MCMCcoal Bayesian estimation of population sizes in a known tree of species.
`http://abacus.gene.ucl.ac.uk/software/MCMCcoal.html`

LDHAT Composite likelihood method for estimating recombination rates.
`http://www.stats.ox.ac.uk/∼mcvean/LDhat/`

Hotspotter Product of Approximate Conditionals likelihood inference of recombination rates.
`http://www.biostat.umn.edu/∼nali/SoftwareListing.html`

Recs Coalescent inference of recombination hotspots.
`http://www.maths.lancs.ac.uk/∼fearnhea/software/Rec.html`

sequenceLD Approximate likelihood inference of recombination rate.
`http://www.maths.lancs.ac.uk/∼fearnhea/software/Rec.html`

sequenceLDhot Approximate likelihood inference of recombination hotspots.
`http://www.maths.lancs.ac.uk/∼fearnhea/`

popgen R package that includes neutral coalescent simulation of samples with recombination.
`http://www.stats.ox.ac.uk/mathgen/software.html`

CodonRecSim Simulation of sequence evolution under a codon model in a coalescent with recombination.
`http://www.binf.ku.dk/~rasmus/webpage/CodonRecSim.html`

SelSim Simulates samples under natural selection.
`http://www.stats.ox.ac.uk/mathgen/software.html`

hap and **dip** Simulate samples at a locus with natural selection.
`http://www.maths.lancs.ac.uk/~fearnhea/software/PS.html`

ms Simulates samples under a neutral coalescent with recombination and mutation.
`http://home.uchicago.edu/~rhudson1/source/mksamples.html`

SIMCOAL Simulates sequence evolution in a coalescent with migration.
`http://cmpg.unibe.ch/software/simcoal/`

Treevolve Simulates sequences evolving on a recombining coalescent with neutral mutation, population growth and migration.
`http://evolve.zoo.ox.ac.uk/software.html?id=treevolve`

Mesquite Can simulate coalescents within species trees.
`http://mesquiteproject.org/Mesquite_Folder/docs/mesquite/popGen/PopGen.html#simulating`

I have not tried to describe which operating systems each program requires. The programs in this list are all free. I have omitted here a number of programs that infer haplotypes rather than model parameters. By the time you read this, there will probably be many more programs. Unfortunately, as yet there is no central list of coalescent programs being maintained on the web.

1.5 The wave of the future

I have introduced the coalescent and some of the major approaches to inference that use it. I could not describe the full range of active work now going on, particularly with models of natural selection, models of recombination hot spots, and reconstruction of haplotypes from diploid data. We have passed the time when a single article could cover coalescent approaches. At least one major book on the coalescent has recently appeared [23]. It concentrates more on the population genetic phenomena than on inference methods.

To many researchers on evolutionary genetics and population genomics coalescent inference methods may appear to be one of the major approaches, but only one. This perception will change, I hope rapidly. Coalescent inference methods are destined to replace most (perhaps all) other inference methods in these fields. They are currently limited by their computational burden, and by the difficulty of developing software to treat all cases. As those limitations are overcome, we will look back on the past decade as the period in which the major methods of analysis of population-level data developed, a period in which molecular evolution and population genetics began their ultimate merger. Students who now see coalescents as one interesting topic among many will ultimately understand that coalescents are the fundamental tool for analyzing evolutionary data near the species level.

Acknowledgments

Work on this paper was supported by NIH grant R01 GM071639. I wish to thank the reviewers for many helpful comments, and for explaining to me what kind of book they would have written instead of this article.

References

[1] Bahlo, M. and Griffiths, R. C. (2000). Inference from gene trees in a subdivided population. *Theoretical Population Biology*, **57**, 79–95.

[2] Beerli, P. B. and Felsenstein, J. (1999). Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics*, **152**, 763–773.

[3] Beerli, P. B. and Felsenstein, J. (2001). Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences, USA*, **98**, 4563–4568.

[4] Crow, J. F. and Kimura, M. (1964). The number of alleles that can be maintained in a finite population. *Genetics*, **49**, 725–738.

[5] De Iorio, M. and Griffiths, R. C. (2004). Importance sampling on coalescent histories. I. *Annals of Applied Probability*, **36**, 417–433.

[6] De Iorio, M. and Griffiths, R. C. (2004). Importance sampling on coalescent histories. II: Subdivided population models. *Annals of Applied Probability*, **36**, 434–444.

[7] Drummond, A. J., Nicholls, G. K., Rodrigo, A. G., and Solomon, W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, **161**, 1307–1320.

[8] Edwards, A. W. F. (1970). Estimation of the branch points of a branching diffusion process. *Journal of the Royal Statistical Society, Series B*, **32**, 155–174.

[9] Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, **3**, 87–112.

[10] Fearnhead, P. and Donnelly, P. (2001). Estimating recombination rates from population genetic data. *Genetics*, **159**, 1299–1318.

[11] Fearnhead, P. and Donnelly, P. (2002). Approximate likelihood methods for estimating local recombination rates. *Journal of the Royal Statistical Society, series B*, **64**, 657–680.

[12] Feller, W. (1951). Diffusion processes in genetics. In *Proc. Second Berkeley Symposium on Mathematical Statistics and Probability* (ed. J. Neyman), pp. 227–246. University of California Press, Berkeley and Los Angeles.

[13] Felsenstein, J. (1992). Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Genetical Research*, **59**, 139–147.

[14] Felsenstein, J. (2006). Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more loci? *Molecular Biology and Evolution*, **23**, 691–700.

[15] Felsenstein, J., Kuhner, M. K., Yamato, J., and Beerli, P. (1999). Likelihoods on coalescents: a Monte Carlo sampling approach to inferring parameters from population samples of molecular data. In *Statistics in Molecular Biology and Genetics* (ed. F. Seillier-Moiseiwitsch), IMS Lecture Notes-Monograph Series, volume 33, pp. 163–185. Institute of Mathematical Statistics and American Mathematical Society, Hayward, California.

[16] Fu, Y. X. and Li, W.-H. (1993). Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Genetics*, **134**, 1261–1270.

[17] Golding, G. B. (1984). The sampling distribution of linkage disequilibrium. *Genetics*, **108**, 257–274.

[18] Griffiths, R. C. (1981). Lines of descent in the diffusion approximation of neutral Wright-Fisher models. *Theoretical Population Biology*, **17**, 37–50.

[19] Griffiths, R. C. (1989). Genealogical-tree probabilities in the infinitely-many-site model. *Journal of Mathematical Biology*, **27**, 667–680.

[20] Griffiths, R. C. and Marjoram, P. (1996). Ancestral inference from samples of DNA sequences with recombination. *Journal of Computational Biology*, **3**, 479–502.

[21] Griffiths, R. C. and Tavaré, S. (1994). Sampling theory for selectively neutral alleles in a varying environment. *Philosophical Transactions: Biological Sciences*, **344**, 403–410.

[22] Griffiths, R. C. and Tavaré, S. (1999). The ages of mutations in gene trees. *Annals of Applied Probability*, **9**, 567–590.

[23] Hein, J., Schierup, M. H., and Wiuf, C. (2005). *Gene Genealogies, Variation and Evolution. A Primer in Coalescent Theory.* Oxford University Press, Oxford.

[24] Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, **23**, 183–201.

[25] Hudson, R. R. (2001). Two-locus sampling distributions and their application. *Genetics*, **159**, 1805–1817.

[26] Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.

[27] Hudson, R. R. and Kaplan, N. L. (1988). The coalescent process in models with selection and recombination. *Genetics*, **120**, 831–840.

[28] Kaplan, N. L., Darden, T., and Hudson, R. R. (1988). The coalescent process in models with selection. *Genetics*, **120**, 819–829.

[29] Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and Their Applications*, **13**, 235–248.

[30] Kingman, J. F. C. (1982). Exchangeability and the evolution of large populations. In *Exchangeability in Probability and Statistics. Proceedings of the International Conference on Exchangeability in Probability and Statistics, Rome, 6th-9th April, 1981, in honour of Professor Bruno de Finetti* (ed. G. Koch and F. Spizzichino), pp. 97–112. North-Holland Elsevier, Amsterdam.

[31] Krone, S. M. and Neuhauser, C. (1997). Ancestral processes with selection. *Theoretical Population Biology*, **51**, 210–237.

[32] Kuhner, M. K., Beerli, P., Yamato, J., and Felsenstein, J. (2000). Usefulness of single nucleotide polymorphism data for estimating population parameters. *Genetics*, **156**, 439–447.

[33] Kuhner, M. K. and Felsenstein, J. (2000). Sampling among haplotype resolutions in a coalescent-based genealogy sampler. *Genetic Epidemiology*, **19 (Supplement 1)**, S15–S21.

[34] Kuhner, M. K., Yamato, J., and Felsenstein, J. (1995). Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics*, **140**, 1421–1430.

[35] Kuhner, M. K., Yamato, J., and Felsenstein, J. (1998). Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics*, **149**, 429–434.

[36] Kuhner, M. K., Yamato, J., and Felsenstein, J. (2000). Maximum likelihood estimation of recombination rates from population data. *Genetics*, **156**, 1393–1401.

[37] Li, N. and Stephens, M. (2003). Modeling linkage disequilibrium and indentifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, **165**, 2213–2233 (Erratum, vol. 167, p. 1039, 2004).

[38] Marjoram, P., Molitor, J., Plagnol, V., and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences, USA*, **100**, 15324–15328.

[39] McVean, G., Awadalla, P., and Fearnhead, P. (2002). A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics*, **160**, 1231–1241.

[40] Neuhauser, C. and Krone, S. M. (1997). The genealogy of samples in models with selection. *Genetics*, **145**, 519–534.

[41] Nielsen, R. (1998). Maximum likelihood estimation of population divergence times and population phylogenies under the infinite sites model. *Theoretical Population Biology*, **53**, 143–151.

[42] Nielsen, R. (2000). Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, **154**, 931–942.

[43] Nielsen, R. and Wakeley, J. (2001). Distinguishing migration from isolation: A Markov Chain Monte Carlo approach. *Genetics*, **158**, 885–896.

[44] Plagnol, V. and Tavaré, S. (2002). Approximate Bayesian Computation and MCMC. In *Monte Carlo and Quasi-Monte Carlo Methods 2000: Proceedings of a Conference held at Hong Kong Baptist University, Hong Kong SAR,*

China, Nov. 27-Dec.1, 2000 (ed. K. T. Fang, F. J. Hickernell, and H. Niederreiter), pp. 99–114. Springer-Verlag, London.

[45] Robertson, A. and Hill, W. G. (1983). Population and quantitative genetics of many linked loci in finite populations. *Proceedings of the Royal Society of London, Series B. Biological Sciences*, **219**, 253–264.

[46] Rodrigo, A. and Felsenstein, J. (1999). Coalescent approaches to HIV-1 population genetics. In *The Evolution of HIV* (ed. K. A. Crandall), pp. 233–272. Johns Hopkins University Press, Baltimore.

[47] Slatkin, M. and Hudson, R. R. (1991). Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics*, **129**, 555–562.

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

[49] Tajima, F. (1983). Evolutionary relationship of DNA-sequences in finite populations. *Genetics*, **105**, 437–460.

[50] Takahata, N. (1988). The coalescent in two partially isolated diffusion populations. *Genetical Research*, **52**, 213–222.

[51] Takahata, N. and Nei, M. (1995). Gene genealogy and variance of interpopulational nucleotide differences. *Genetics*, **110**, 325–344.

[52] Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, **7**, 256–276.

[53] Weiss, G. and von Haeseler, A. (1998). Inference of population history using a likelihood approach. *Genetics*, **149**, 1539–1546.

[54] Wilson, I. J. and Balding, D. J. (1998). Genealogical inference from microsatellite data. *Genetics*, **50**, 499–510.

[55] Wilson, I. J., Weale, M. E., and Balding, D. J. (2003). Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **166**, 155–201.

[56] Wiuf, C. and Hein, J. (1999). Recombination as a point process along sequences. *Theoretical Population Biology*, **55**, 248–289.