Coalescing into the 21st Century: An Overview and Prospects of Coalescent Theory

Yun-Xun Fu
Human Genetics Center, SPH, University of Texas, P.O. Box 20334, Houston, Texas 77225

and

Wen-Hsiung Li
Ecology and Evolution, University of Chicago, 1101 East 57th Street, Chicago, Illinois 60637

Received July 16, 1998

1. INTRODUCTION

A central theme in population genetics is understanding how a population evolves under a given set of conditions. Evolution is a forward process in the sense that a population changes its characteristics with time. Consequently, prospective approaches in population genetics dominated the field in the past. The focus of a prospective approach is usually a certain characteristic or quantity, such as the frequency of a mutant gene in the population. However, empirically the characteristics of a natural population are usually examined by taking samples from the population. Interesting biological questions that arise from a sample are mostly retrospective, such as the history of the population that gave rise to the sample, or the evolutionary mechanisms responsible for the characteristics observed. The rapid accumulation of DNA sequence data since the 1980s has transformed the mainstream of population genetics research from prospective to retrospective, from demonstration of principles to inference of events that happened in the past.

To infer the past from a sample taken from a present population, a new approach is required. Coalescent theory arose from this necessity. The essence of coalescent theory is to start with a sample, and trace backward in time to identify events that occurred in the past since the most recent common ancestor of the sample. Since the seminal work of Kingman (1982a, b), coalescent theory has been the most active topic in theoretical population genetics, and it is now widely recognized as the cornerstone for various statistical analyses of molecular population samples. The usefulness of the theory comes mainly from three features. First, it is a sample-based theory. Since the study of a population usually relies on a sample of individuals from that population, a theory that describes the properties of a sample is more relevant than the classical population genetics theory that describes the properties of the entire population. Second, it is a highly efficient approach. An important by-product of coalescent theory is the development of highly efficient algorithms for simulating population samples under various population genetics models, allowing various aspects of a model to be examined numerically. Third, coalescent theory is particularly suitable for molecular data, such as DNA sequence samples, which contain rich information about the ancestral relationships among the individuals sampled.

Various aspects of coalescent theory have been reviewed previously (Tavaré, 1984; Takahata, 1991;
2. COALESCENT PROCESSES AND RECENT PROGRESS

Consider a sample of \( n \) sequences of a DNA region from a population and assume that there is no recombination between sequences. Then the \( n \) sequences are connected by a single phylogenetic tree or genealogy, in which the root of the tree is the most recent common ancestor (MRCA) of these \( n \) sequences. The history of the \( n \) sequences can be viewed from two different perspectives. If one starts with the MRCA and looks forward in time, one sees that once in a while one of the existing sequences splits into two and along the way mutations accumulate. The impression from this prospective view is the divergence of sequences. On the other hand, if one starts with the sample of sequences and looks backward in time, one sees that the number of ancestral sequences becomes fewer and fewer and the sequences become more and more similar. This retrospective view gives the impression of coalescence.

It is convenient to consider coalescent theory as consisting of two inter-related components. The first one is the stochastic processes leading to the MRCA from a sample under various population genetics models. The second one is the statistical properties of and inference methods based on observable quantities in a sample, which can be summary statistics or more detailed patterns of polymorphism. The coalescent process of two sequences from a population (of diploid organisms) under the neutral Wright–Fisher model (see the next section for definition) is as follows. Looking backward in time, the probability of coalescence at the previous generation, i.e., the two sequences in the current generation came from a single ancestral sequence in the previous generation, is 1/(2\( N \)), where \( N \) is the effective population size. The probability that coalescence occurred \( t+1 \) generations ago is given by the geometric distribution

\[
\frac{1}{2N} \left( 1 - \frac{1}{2N} \right)^t,
\]

which can be approximated by an exponential distribution with mean equal to 2\( N \). Assume that the number of mutations that occurred on a sequence in a given time period is a Poisson variable. Then the mean time of 2\( N \) generations separating the two sequences implies that the mean number of mutations in the two sequences is \( \theta = 4N\mu \), where \( \mu \) is the mutation rate per sequence per generation.

An important by-product of coalescent theory is very efficient algorithms for simulating an evolutionary process and therefore population samples from that process. Statistical properties of observable quantities in a sample are necessary for making a proper inference about the history of the population and the mechanism responsible for the pattern of polymorphism.

2.1. Population Genetics Models

To date, coalescent processes have been studied for a number of population models. However, the most extensively studied model is the so-called neutral Wright–Fisher model, which assumes that the sequences in a population at a generation are a random sample (with replacement) from those in the previous generation and that all mutations at the locus under question are selectively neutral. Models with recombination have been the subject of several recent studies (Griffiths and Marjoram, 1996, 1997; Hey and Wakeley, 1997) while earlier studies on recombination were reviewed by Hudson (1991). Recombinations break linkage between different regions of a chromosome or nucleotide sites and consequently generate a number of correlated genealogies. Concise representation of such correlated genealogies using graph theory is one contribution of Griffiths and Marjoram (1996, 1997), in addition to their method for estimating the number of recombination events in a sample.

Models with natural selection have also received considerable attention. Earlier studies were also reviewed by Hudson (1991). More recent studies include background selection (Hudson and Kaplan, 1994; Charlesworth et al., 1995) and haploid selection (Krone and Neuhauser, 1997; Neuhauser and Krone, 1997). The ancestral selection graph approach by Krone and Neuhauser (1997) appears to be promising. Although their analysis is based on the Moran model, which is strictly a haploid model, Neuhauser and Krone (1997) argued that their ancestral selection graph is also an approximation to the diploid selection model because the limiting processes of both models are the same.

Until recently the assumption of a randomly mating population (or local population) has been made in most studies. For organisms with high mobility such as many animal populations, the assumption is a reasonable approximation. However, this assumption is not adequate...
Overview and Prospects of Coalescent Theory

for many plant populations whose reproduction is through selfing or partially selfing. With an increasing number of plant population studies at the molecular level, the usefulness of the coalescent approach for partially selfing populations is increasingly being recognized (Slatkin, 1991; Milligan, 1996; Clegg, 1997; Nordborg and Donnelly, 1997; Fu, 1997a). The general coalescent framework for partially selfing populations was recently developed by Nordborg and Donnelly (1997) and Fu (1997a).

2.2. Polymorphism in a Sample

Under the neutral Wright–Fisher model, the time between two successive coalescent events, the so-called coalescent time, is approximately an exponential variable with mean $4N/k(k - 1)$, where $k$ is the number of ancestral sequences of the sample between the two coalescent events. Under more complex models such as recombination and population subdivision, a coalescent time is often a sum of exponential variables. Usually the number of mutations occurring in a sequence during a given time period is assumed to be a Poisson variable. A complete specification of the coalescent process and the mutation process allows the pattern of polymorphism in a sample to be characterized.

What characteristics of a sample should be used for summarizing sample polymorphism and for statistical inference is of great importance from both theoretical and practical points of view. Until recently, most studies have focused on two summary statistics. One is the number $K$ of mutations (which is the number of segregating sites in a sample under the assumption of the infinite-site model), and the other is the mean number $\pi$ of nucleotide differences between two sequences in a sample. Under the neutral Wright–Fisher model, $K$ is the sum of $n - 1$ geometric variables, whose distribution has been obtained analytically. The mean and variance of $K$ are particularly well known due to Watterson (1975). The distribution of $\pi$ is unknown but its mean and variance were derived by Tajima (1983). The statistical properties of $K$ in a subdivided population were studied by Tajima (1989a) and Wakeley (1998).

Although $K$ and $\pi$ have a number of important applications and are likely to continue to play important roles in population genetics, they convey only a fraction of the information available in a sample. An alternative approach is to make statistical inferences based on a more detailed pattern of polymorphism or even on the raw data. Mutations in a sample can be viewed on a finer scale. A mutation which is inherited by $i$ sequences in a sample is said to be of size $i$. Therefore, mutations in a sample fall into $n - 1$ different sizes. As both $K$ and $\pi$ can be computed from the frequencies of mutations of different sizes, the latter contain more information than the former. How much more information is captured by such finer grouping for a population model remains to be studied, but Fu (1994b) showed that a least-squares estimator of $\theta = 4N\mu$ based on these frequencies is very close in terms of accuracy to the best possible estimator. Fu (1995) has obtained the means and variances of the numbers of mutations of various sizes and their covariances. Another classification of a mutation can be made according to the frequencies of the two segregating nucleotides at the site. The statistical properties of this classification can be obtained from those of the numbers of mutations of various sizes (Fu, 1995).

2.3. Statistical Inference

Since coalescent theory provides a powerful framework for analyzing polymorphism data, which have dramatically increased in volume in the past decade, it is not surprising that the development of coalescent-based statistical methods for analyzing DNA sequence samples has received much attention in recent years.

Estimation of $\theta$. Since $\theta = 4N\mu$ is probably the most important population parameter in the stochastic theory of population genetics, methods for estimating $\theta$ are of special importance. Two widely used estimators are $K = K/a_n$ (Watterson, 1975) and $\bar{\pi} = \pi$ (Tajima, 1983), where $a_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n - 1}$. Both estimators are unbiased under the neutral Wright–Fisher model but Watterson’s estimator has a smaller variance than Tajima’s. Since both estimators make little use of the genealogical relationships among the sequences in a sample, incorporating this information into the estimation was the focus of several new methods. First, Felsenstein (1992) and Fu and Li (1993a) showed that there is considerable room for improvement. In particular, Fu and Li (1993a) derived the minimum variance $V_{\text{min}}$ of any unbiased (or nearly unbiased) estimator of $\theta$, and showed that $V_{\text{min}}$ can be substantially smaller than that of Watterson’s estimator, although the latter is approaching an optimal estimator with increasing sample size. Fu (1994a) then demonstrated that a nearly unbiased phylogenetic estimator (UPBLUE) of $\theta$ achieved a variance almost as small as $V_{\text{min}}$ and therefore can be considerably smaller than that of Watterson’s estimator. Fu’s method was based on a generalized least-squares method, taking advantage of the linear relationship between the expected number of mutations on a branch and $\theta$. Alternatively, the maximum-likelihood approach can be used. To date, three different likelihood methods
have been developed: (1) Griffiths and Tavaré (1994a, b, c, 1995) developed a Monte Carlo method by taking advantage of earlier work (Griffiths, 1989) on a recurrence equation for the probability of a polymorphism; (2) Kuhner et al. (1995) developed another Monte Carlo estimator using the Metropolis–Hastings method; and (3) Fu (1998) proposed a maximum-likelihood method for use with the frequencies of mutations of various sizes.

**Tests of Selective Neutrality.** The neutral theory of molecular evolution (Kimura, 1968, 1983) postulates that the majority of mutations that contribute significantly to the genetic variation in natural populations are neutral or nearly neutral. The neutral theory has profound influences on molecular population genetics and there have been much interest in testing the theory over the years. The rapid accumulation of DNA sequence polymorphism data provides many opportunities for testing various aspects of the theory. Instead of testing the neutral mutation hypothesis, one may test the hypothesis of strict neutrality, which postulates that the locus in question evolves according to the Wright–Fisher model and that all mutations are selectively neutral. Reasoning that Watterson’s estimator and Tajima’s estimator of $\theta$ are affected by natural selection to different extents, Tajima (1989b) developed a test of neutrality by using the difference between the two estimators. Fu and Li (1993b) proposed several tests by utilizing mutations occurring on internal and external branches of the genealogy of a sample, and Fu (1996b) developed two other tests making use of the pattern of segregating sites. Neutrality tests can also be developed by utilizing Ewens’ (1972) sampling theory. One such test by Fu (1997b) appears to be very powerful for detecting the presence of genetic hitchhiking when there is no recombination.

Several studies have been conducted to evaluate the powers of these tests (Braverman et al., 1995; Simonsen et al., 1995; Fu, 1996b, 1997b). It should be noted that the tests discussed here are so-called pure significance tests in statistics, because they are designed to test whether an observed polymorphism is consistent with the neutral model of evolution, without taking into consideration any alternative model (although the potential effect of alternative models was important in designing these tests). Since the strict neutral model of evolution is usually the starting assumption for data analysis, and since there exists a large number of alternative models, pure significance tests are often the desirable form of test. However, when the consideration of a specific alternative model is justified, more powerful tests based on the likelihood ratio can potentially be developed (e.g., Griffiths and Tavaré, 1994b, for testing constant population size versus exponential growth). Because models of a very different nature, such as genetic hitchhiking and population growth, can give rise to very similar patterns of polymorphism (e.g., Fu, 1997b), caution must be exercised to avoid over-confidence on the specific alternative model when the strict neutral model is rejected. Also the tests mentioned use only intra-specific polymorphism. Therefore, they test if an observed polymorphism is consistent with the strict neutral model within a short period of $4N$ generations on average. A combination of both intra- and inter-specific polymorphism data provides an opportunity to test the neutral hypothesis over a longer period. One such test was the McDonald–Kreitman test (McDonald and Kreitman, 1991; see Li, 1997, for more discussion).

**Estimation of the Age of the Common Ancestor.** Estimation of the age $T$ of the MRCA of a sample is another hot area of research, largely due to its relevance to the origin of modern humans. The theoretical development was accelerated by a debate over how to estimate $T$ for a sample of 37 Y-chromosome sequences by Dorit et al. (1995) in which there was no variation (Fu and Li, 1996; Donnelly et al., 1996; Weiss and von Haeseler, 1996). Estimation of $T$ can now be made based on (1) the number $K$ of segregating sites (Fu, 1996a; Tavaré et al., 1997), (2) the maximum number of nucleotide differences between two sequences in a sample (Fu and Li, 1997), (3) $K$ and the number of alleles (Griffiths and Tavaré, 1996), and (4) the full set of polymorphisms (Griffiths and Tavaré, 1994a). It is interesting to note that all these estimators are Bayesian estimators. A likelihood estimator of $T$, say, based on $K$, can be devised, but it tends to give too small an estimate when $K$ is small and too large an estimate when $K$ is large due to ignoring prior information on $T$ (Fu, 1996a). Therefore, the maximum-likelihood method does not seem to be appropriate for age estimation. Another important age estimation problem is the divergence time between two populations; an interesting coalescent estimator was developed by Nielsen (1998) extending Griffiths and Tavaré’s approach.

**Estimation of Recombination Rate.** The recombination rate $r$ is another important parameter in population genetics. Unlike mutations which give rise to segregating sites (most often different segregating sites), recombinations very often do not leave a trace. Therefore, it is more difficult to estimate this parameter than $\theta$. Hey and Wakeley (1997) proposed an estimator which in essence is similar to that by Hudson and Kaplan (1985) but performs better when the sample size is small. The maximum-likelihood approach is likely to be more powerful for estimating $r$. One maximum-likelihood method was
developed by Griffiths and Marjoram (1996), but the properties of this estimator remain to be studied.

Other progress includes estimation of the ancestral population size (Takahata et al., 1995; Wakeley and Hey 1997), estimation of the migration rate (Slatkin and Maddison, 1989; Hudson and Kaplan, 1994), estimation of the selfing rate (Milligan, 1996; Nordberg and Donnelly, 1997; Fu, 1997a), and estimation of the parameters of the step-wise mutation model (Nielsen, 1997; Fu and Chakraborty, 1998; Wilson and Balding, 1998).

3. SOME OUTSTANDING ISSUES AND PROSPECTS OF COALESCENT THEORY

There are many aspects of coalescent theory that remain to be studied, some of which are briefly discussed below. Our selection of issues is subjective and reflects in part our experience. Nevertheless, we hope that this discussion may stimulate more critical thinking and further research.

The framework on which powerful statistical inferences can be developed is what makes coalescent theory successful. However, although a novel theory can inspire and guide data collection, more often the development of statistical methods is stimulated by available data and biological questions. So, before discussing the prospects of coalescent theory, it is necessary to foresee what kind of molecular data will be most abundant or will be collected in the near future.

3.1. Molecular Data

To date, most applications of coalescent theory have been to samples from the control region of human mitochondrial DNA (mtDNA). Because mtDNA is haploid without recombination, DNA polymorphism is available naturally in haplotypes. However, since mtDNA represents only a single locus, its usefulness for studying human population history is limited. Recent data from the non-recombining region of the human Y chromosome suffer the same drawback. For a better picture of the human history, more loci must be studied, which means that the nuclear genome is the future battleground. The study on the β-globin gene tree by Harding et al. (1997) represents a recent notable example of a human population study using a nuclear locus.

Two kinds of polymorphism are common. One is segregating sites, which are now fashionably termed single nucleotide polymorphisms (SNPs), and the other is micro-satellite data. The latter have been used extensively as genetic markers and for forensic purposes, and also can be used for population studies. Unfortunately, because of extremely high mutation rates, and unknown mutation mechanisms, ancestral information from micro-satellite loci is often ambiguous. Therefore, micro-satellite data are less ideal for evolutionary studies, and consequently their importance is likely to diminish gradually.

It has been estimated that the nucleotide diversity of the human nuclear genome is about 0.1% (Li and Sadler, 1991). For two randomly selected sequences, this number translates into one polymorphic site for every 1000 nucleotides. For a large sample of sequences, one polymorphic site is expected for every 200–500 nucleotides. Since the human nuclear genome contains about three billion nucleotides, several million polymorphic sites are expected to exist. Although the number of known polymorphic sites is not large at present, the situation will likely change dramatically in the near future with the progress of the Human Genome Project and the recent initiative, by the National Institutes of Health of the United States, to catalog a large number of SNPs. Although the initiative is aimed at finding disease-causing genes, a large impact on population genetics by the availability of a large collection of SNPs is inevitable because a huge volume of data can be generated economically by using them to genotype population samples. For each SNP, a population screening results in the number of sequences carrying each segregating nucleotide, and screening a large number of SNPs results in a pattern of segregating sites. At present and in the foreseeable future, polymerase chain reaction (PCR) is the method of choice for amplifying DNA segments for detecting polymorphism. Since segments on both chromosomes are amplified, one can usually determine the genotype of each polymorphic site for an individual. Determination of haplotype sequences for a large segment requires further experiments, which can be assisted to some degree by statistical inferences (e.g., Clark, 1990). Overall the process is time-consuming and costly. Such experiments usually generate the pattern of segregating sites first and then gradually resolve the haplotypes. Haplotype determination is even more difficult for a polyploid genome, such as many plants.

Among various forms of polymorphism information, complete haplotype sequences have the highest resolution. If cost is of no concern, then obtaining haplotype sequences of a large sample from many different regions of the genome would be a good strategy. Otherwise, one needs to explore whether other forms of polymorphism are more effective than haplotype sequences for a given budget. The most important alternative to haplotype
sequences is the pattern of segregating sites or array of SNPs. Obviously, whether a particular form of data will be collected depends on not only the cost and biological questions addressed, but also the availability of theory and methods for analyzing such data.

### 3.2. Population Genetics Models

The neutral Wright–Fisher model has played and will continue to play a central role in coalescent theory, not only because it is easier to obtain analytical results under the model, but also because it provides the base for comparing various models. The coalescent processes of many alternative models have been studied but in general our knowledge of an alternative model, such as natural selection, is far less than that of the neutral Wright–Fisher model. For many natural populations, the neutral Wright–Fisher model is a very rough approximation at best. Take human populations, for example. A realistic model should incorporate migration and population growth, and also needs to include recombination and natural selection depending on the loci studied. Although coalescent processes involving multiple evolutionary forces have been studied (see, for example, Hudson, 1991), the interplay among and relative importance of various forces are not well understood. One problem is that a pattern of polymorphism may be generated with more or less the same likelihood by a number of different models. For example, population growth and genetic hitchhiking both result in a similar pattern of excess of rare alleles (e.g., Fu, 1997b). For any given sample from a population, it is a great challenge to identify with confidence the mechanisms responsible for the observed pattern of polymorphism. For this purpose, we need to have a good understanding of the effect of various evolutionary forces.

Until recently, almost all the models studied in coalescent theory assumed random mating among individuals in a population (at least in each local population). Non-random mating in natural populations is not uncommon. For example, many plant populations undergo both out-crossing and self-fertilization and many animal populations (such as primates) have well-recognized social structures. Therefore, it will be useful to develop coalescent theory incorporating various types of non-random mating. Although some of the complexities due to non-random mating may be circumvented by using the long-term effective population size, such an approach is unlikely to be satisfactory, especially when the dynamics of the size of a population of interest.

Coalescent theory has always been developed for a population with a finite size. Bacterial or viral populations are usually studied using deterministic models because of their huge population sizes. However, HIV and some other emerging infectious disease agents may provide interesting applications of coalescent theory. The importance of random genetic drift has been highlighted by a recent study (Leigh Brown and Richman, 1997) showing that the effective size of a host HIV population is rather small. Coalescent theory should be able to contribute significantly to the genetic analyses of HIV and other pathogenic populations. Of course, there are some specificities in such populations and both within-host and between-host dynamics need to be considered. Furthermore, longitudinal samples are common in studying epidemics, so both coalescent theory and statistical methods need to be developed for analyzing longitudinal samples. One interesting application of a longitudinal sample is to estimate the length of a generation of HIV (Rodrigo et al, 1999).

Another area where the development of coalescent theory is lacking is the evolutionary analysis of quantitative trait loci (QTL), despite a recent surge of interest in locating and studying the molecular basis of quantitative traits. Coalescent theory should be useful for answering a number of interesting questions, such as the trait value of the most recent common ancestor and the evolutionary mechanisms for a QTL. It appears that the key to a coalescent analysis of QTL is the specification of the relationship between a molecular polymorphism and the value of a trait. Many traits of interest are probably subject to natural selection; it would be a challenge to develop a coalescent theory for a trait which is only partly genetically controlled. In the community of gene hunters, searching for a particular allele associated with a disease (or a trait) in a random sample of patients and normal individuals is increasingly being considered the best hope for identifying genes for a complex trait, that is, a trait that does not follow a simple Mendelian ratio of transmission (e.g., Collins et al., 1997; Schaefer and Hawkins, 1998). Coalescent theory may prove to be valuable in such efforts.

It should be emphasized that coalescent theory as a whole is an approximation to the evolutionary process of a population. When the population size is large relative to sample size, coalescent theory should be quite accurate, but little can be found in the literature on the accuracy of coalescent approximation under various combinations of population size, sample size, mutation rate, etc. Such numerical studies are useful in understanding the limit of the coalescent approach, and are particularly important for more complex models, such as models with selection, because analytical analysis of the accuracy of approximation is more difficult in complex models.
3.3. Statistical Inference

Statistical inference may include parameter estimation and hypothesis testing. Various inference methods have been developed in the 1990s. The field has witnessed a transition from simple moment estimators to various more sophisticated methods based on principles such as generalized least squares, maximum likelihood, and Bayesian. This rapid expansion in methodology will likely continue for a while, but we have reached a point where the properties, such as bias, variance, and robustness of some important estimators, need to be investigated, and also comparative studies of methods have become important. An example is the methods based on Monte Carlo Markov chain (MCMC). Such methods are sound in principle, but whether a particular method is good depends on how well the computational implementation has been made. Because MCMC methods generally demand large computer resources, an examination of the properties of these methods is even more demanding in computer speed. Consequently, the properties of most estimators based on MCMC have not yet been well studied.

Haplotypes vs Segregating Sites. Haplotype sequence data have been used for population and evolutionary studies for a while, so the development of inference methods based on haplotype sequences is likely to be a major focus for the foreseeable future. Monte Carlo-based methods for haplotype sequence data are very promising but more efficient algorithms and implementations will be needed to effectively study complex models, involving migration, recombination, selection, and population growth. For some applications, least-squares methods can be more effective.

With the anticipated rapid accumulation of SNPs in the near future, statistical methods based on segregating sites will become important. The Human Genome Project and related projects will undoubtedly accelerate the development of such methods. One fundamental question, however, is whether data in the form of patterns of segregating sites are sufficiently informative. Fu (1994b) showed that estimation of the age of the MRCA of a sample based on the pattern of segregating sites may be of quality similar to that based on complete haplotype sequences. Although further studies are required for a better understanding of the relative strength and weakness of patterns of segregating sites, it is clear that polymorphism of this form is rich in information content. To date, only a few methods (e.g., Tajima, 1989b; Fu and Li, 1993b; Fu, 1994b, 1995, 1996b, 1997b, 1998) utilize patterns of segregating sites.

Computational Techniques. One of the most powerful ways to make inference is through computing the probability of the characteristics of a sample. There are three ways a probability can be computed. First, an analytical formula is available. This is usually the most preferred situation. Second, find the numerical solution for the equations that specify the probability. Often the numerical value of a probability is what is needed for making inferences. The third is estimation. When the first two approaches fail, one has to rely on the estimation of a probability. An analytical form for the distribution of patterns of polymorphism is only occasionally found. One example is Ewens’ (1972) sampling formula. Another is the distribution of the number of segregating sites (Tavaré, 1984). When the pattern of polymorphism is complex, there is little chance of finding an analytical form of its distribution that is simple enough for a direct computation. However, because a coalescent process can be formulated as a Markov chain, the probability of a pattern of polymorphism can often be derived as a recurrence equation. Consider the number of segregating sites for example. Let $p_s(K)$ be the probability of the number of segregating sites in a sample of $n$ sequences. Looking backward in time, the first event is either a coalescent or a mutation, with probabilities $(n-1)/(\theta+n-1)$ and $\theta/(\theta+n-1)$, respectively. This leads to the following recurrence equation:

$$p_s(K) = \frac{n-1}{\theta+n-1} p_s(K-1) + \frac{\theta}{\theta+n-1} p_s(K-1).$$

When a recurrence equation is relatively simple, such as that given above, the value of the probability can be solved by iteration procedure, which has been done in two different ways by Griffiths (1989) and Fu (1998) for a modest sample size and a modest amount of polymorphism. In general one has to rely on a numerical estimation of the probability.

One effective way to estimate the probability of an observed pattern of polymorphism is to use simulated samples histories through a recurrence equation. This approach has been used successfully by Griffiths and his colleagues (e.g., Griffiths and Tavaré, 1994a-c, 1995; Griffiths and Marjoram, 1996). It, however, requires
knowledge of the haplotype sequences. Although recurrence equations for a pattern of segregating sites can be derived, they are in general more complex than those for the case of haplotype sequences, and it is not clear whether a similar computational procedure can be applied to those recurrence equations effectively. One way to circumvent the problem is to estimate the haplotype sequences first by, for example, Clark’s (1990) algorithm, and then to estimate the probability of observing the estimated haplotypes. Further investigation is needed to determine how accurate the estimation of haplotypes is and how errors in the estimated haplotypes are propagated.

An alternative approach for computing the probability of a pattern of segregating sites is to separate the process of realizing a sample into two components. The first is to produce the topology of the genealogy and the second is to generate the mutations on the branches of the genealogy (e.g., Hudson, 1985). In fact, this is the most common way to simulate samples using a coalescent algorithm. Let us examine how the probability can be computed. Let $s = \{z_1, \ldots, z_{n-1}\}$ be the number of mutations of size $i$ and $g = s$ be the number of segregating sites. If it does, proceed to the next stage; otherwise, reject the coalescent and randomly select another pair of sequences and retest until a suitable pair of sequences is found. The algorithm is efficient because it is necessary to start checking only when there are relatively few ancestral sequences left in the coalescent process. The new algorithm not only provides a bridge to maximally utilize information in the form of patterns of segregating sites, it can also incorporate partial or complete haplotype sequence information. In the latter case, the probability being computed will be equivalent to that advocated by Griffiths and his colleagues.

Another alternative (Felsenstein, personal communication) is to treat haplotype sequences as missing data, and extend the MCMC method by Kuhner et al. (1995) to simultaneously estimate the likelihoods of the sample polymorphism and the haplotype sequences. This approach will likely demand a huge computer resource.

**Ascertainment Bias.** Another important issue on SNPs is ascertainment bias. Because of the huge number of loci in the genome of a typical higher organism, most SNPs are likely to be discovered through a relatively small sample of individuals. The strategy, which may be adequate for certain purposes, introduces bias in such a way that SNPs with either a high or a low frequency for the mutant nucleotide are less likely to be found. To utilize SNPs for a population and evolutionary study, as well as a study in human medical genetics, proper statistical methods must be developed to correct for ascertainment bias. Since computing the probability of a pattern of segregating sites with ascertainment bias is much more difficult, approaches which require only means, variances and covariances of mutations of various categories such as the least-squares method are more practical.

**Other Remarks.** Finally a few words of caution. With enormous interest in developing sophisticated methods, it is easy for one to lose sight of the merits of those simple but effective methods. Some summary statistics, such as the number of segregating sites and the mean number of nucleotide differences between two sequences, will continue to play an important role in helping us to understand the mechanisms responsible for the observed

$$P(\xi) = \sum_g P(\xi | g) P(g),$$

where $g$ represents a genealogy of $n$ sequences without specifying the time lengths of branches and mutations, and $P(g)$ is the probability of genealogy $g$. The summation is taken over all possible genealogies. For even a modest sample size, the number of possible genealogies is huge, so examining all possible genealogies is out of the question. On the other hand, it is not necessary to do so because genealogies with extremely small $P(\xi | g)$ do not contribute significantly to $P(\xi)$ and therefore need not be examined. The probability of a random genealogy having a non-negligible value of $P(\xi | g)$ can be one in a billion or even smaller, so randomly examining genealogies cannot be a successful strategy in general. Suppose that we can randomly sample genealogies for which the value of $P(\xi | g)$ is not negligible. Then $P(\xi)$ can be estimated by

$$P(\xi) = \frac{P(S)}{L} \sum_{i=1}^L P(\xi | g_i),$$

where $P(S)$ is the probability that a random genealogy has a non-negligible value of $P(\xi | g)$, and $L$ is the number of genealogies sampled. Essential for such an estimator to be practical is an efficient algorithm for generating genealogies with a non-negligible value of $P(\xi | g)$. Such an algorithm is being developed which also allows $P(S)$ to be estimated (Fu, 1999). The idea of the algorithm is as follows: A sample with a mutation of size $i$ implies that the sample genealogy must have a branch of size $i$. While a genealogy is being simulated in a typical fashion starting with $n$ sequences, one can determine whether a coalescent between two sequences can lead to a genealogy that is compatible with the given pattern of segregating sites. If it does, proceed to the next stage; otherwise, reject the coalescent and randomly select another pair of sequences and retest until a suitable pair of sequences is found. The algorithm is efficient because it is necessary to start checking only when there are relatively few ancestral sequences left in the coalescent process. The new algorithm not only provides a bridge to maximally utilize information in the form of patterns of segregating sites, it can also incorporate partial or complete haplotype sequence information. In the latter case, the probability being computed will be equivalent to that advocated by Griffiths and his colleagues.

Another alternative (Felsenstein, personal communication) is to treat haplotype sequences as missing data, and extend the MCMC method by Kuhner et al. (1995) to simultaneously estimate the likelihoods of the sample polymorphism and the haplotype sequences. This approach will likely demand a huge computer resource.
polymorphisms, not only because they are biologically meaningful quantities, but also because they allow a direct comparison of the amount of polymorphism between different samples. In addition to the appeal of simplicity and ease of computation, methods based on summary statistics can be as effective as sophisticated methods in many situations, particularly when the amount of polymorphism at each locus being examined is low.

3.4. Computer Software

Many of the inference methods based on coalescent theory are computation intensive. Although the computer program for carrying out a certain analysis can often be obtained from the author(s), there is increasingly a need to have user-friendly computer packages integrating many useful inference methods. The lack of such software is not surprising because most of the sophisticated inference methods have only recently been developed.

On the other hand, the most important by-product of coalescent theory is efficient algorithms for simulating samples under various models. Much truth can be learned by examining simulated samples and some problems can only be treated this way. Therefore, user-friendly software that allows one to visualize various aspects of a coalescent process can be of great help in research, teaching, and learning coalescent theory.

Some recently developed computer technologies will speed up the development of user-friendly software. For example, the World Wide Web greatly simplifies both the access and the distribution of computer programs, and the new object-oriented programming language Java has the potential to substantially shorten the development time of a software package, to produce highly user-friendly interfaces and output, and to make the program run on all major computer platforms.

ACKNOWLEDGMENTS

This work was supported by NIH Grants R29 GM50428 (Y.X.F), R01 HG01808 (Y.X.F and W.H.L), R01 GM55759 (W.H.L), and R01 GM36998 (W.H.L) and NSF Grant DEB-9707567 (Y.X.F).

REFERENCES


Fu, Y. X. 1996a. Estimating the age of the common ancestor of a DNA sample using the number of segregating sites, Genetics 144, 829–838.

Fu, Y. X. 1996b. New statistical tests of neutrality for DNA samples from a population, Genetics 143, 557–570.


Fu, Y. X. 1997b. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection, Genetics 146, 915–925.


