Coalescent size versus coalescent time with strong selection

R. B. Campbell
Department of Mathematics
University of Northern Iowa
Cedar Falls IA 50614-0506
campbell@math.uni.edu
http://www.math.uni.edu/∼campbell
(319) 273-2447

Running head: Coalescent with strong selection
Keywords: Coalescent, Fixation time, Population genetics, Selection, Variation

ABSTRACT

The coalescent structure (topology) is robust, but the rate of coalescence changes in the presence of selection. The change in the rate of coalescence is not uniform, rather the acceleration of growth of the coalescent is greatest near the common ancestor with little change near fixation. This provides that the reduction in the cumulative size of the coalescent (hence genetic variation) due to selection is much less than the reduction in the coalescent time. If $Ns >> 1$, the coalescent and fixation times are approximately equal to $\frac{\ln 2Ns}{s}$, which is much less than $2N$ which would ensue from neutrtal drift (85% less for $Ns = 10$ and 97% less for $Ns = 100$). However, for those values of Ns, and N ranging between 10^3 and 10^6, the cumulative size of the coalescent is only reduced by 20% to 70% from the neutral case. The reduction in the coalescent time for two alleles versus the neutral case is slightly less than the reduction in the coalescent time for the entire population, approximately 80% and 96% for Ns equals 10 and 100, respectively.

Introduction

The introduction of diffusion methods to population genetics provided the classic result that the time until fixation of a neutral mutation is $2N$ generations (Kimura and Ohta 1969). The coalescent of Kingman (1982a,b) provided the same result for the coalescent time in a population without selection. Indeed, under mild steady state assumptions allowing selection and fluctuation in population size, the coalescent and fixation times are equal (Campbell 1999). But the equations governing the coalescent and fixation times are not amenable to solution in the presence of selection, and only limited results are available.

Neuhauser and Krone (1997) and Krone and Neuhauser (1997) developed a generalization of the coalescent with which they showed that under weak selection the coalescent time is the same as in the neutral case. Van Herwaarden and van der Wal (2002) used an asymptotic approximation to the diffusion equation to obtain fixation times when selection is strong. Both results are based on limits as N (the population size) goes to infinity. We require Ns (s is the Malthusian parameter) to be large in order to ignore some terms in the solution and we generally assume $s < 0.1$. The scope of the coalescent in the neutral case can be found in Donnelly and Tavaré (1995). There are results on the effect of selection at other loci on the coalescent (Barton and Etheridge 2004; Campbell 1999; Kaplan, Darden, and Hudson 1988); but this is not the subject of the present work which considers only one locus (or region with zero recombination, selection acts at the locus of which the coalescent is being studied). Results for the coalescent with fluctuating population size (Kaj and Krone 2003; Polanski, Bobrowski, and Kimmel 2003; Sano, Shimizu, and Iizuka 2004; Griffiths and Tavaré 1994) are far more general than the needs of the present results. In particular, our model reduces to the coalescent of an exponentially growing population (Slatkin and Hudson 1991) of the class of selected alleles (Slatkin 1996). We are only studying the coalescent time prior to fixation of a mutation; the time since fixation (based on other loci) has been studied by Przeworski (2003).

We employ the result that both the coalescent and fixation time are approximately $\frac{\ln 2Ns}{s}$ under strong selection ($Ns >> 1$) (van Herwaarden and van der Wal 2002), for which we provide a different derivation (our result differs from their result in the terms which are ignored). The approximation $\frac{\ln 2Ns}{s}$ can also be obtained heuristically using the observation that drift governs the dynamics of a selected allele when it is rare, but deterministic growth governs the dynamics when it is frequent (Ewens 1979). We are not aware of any published estimates using this heuristic argument, but we obtained the estimate $\frac{\ln 2Ns}{s}$ using the time (approximately $\frac{1}{2s}$ generations) when the frequency of the beneficial mutaion is $\frac{1}{s}$ to demarcate the drift from the deterministic phase. The fixation time $\frac{\ln 2Ns}{s}$ entails a reduction in the coalescent/fixation time by a factor of $\frac{\ln 2Ns}{2Ns}$ from neutrality.

Our derivation considers only the selected mutation as the population, determines that population size by deterministic growth, and uses that population size to rescale the coalescent process. Numerical evaluations of the total number of individuals in the coalescent are obtained for several values of N and s, which manifest a reduction between 20% and 70% from neutrality. (The total number of individuals in the coalescent can be used to estimate the number of alleles segregating since the total number of individuals is the number of opportunities for mutation to occur. The number of alleles segregating is an alternative to heterozygosity as a measure of genetic variation.)

The Model

The model assumes a single beneficial mutation with constant absolute viability $1 + s$ (until fixation). Subsequent mutations share the viability of the allele they replace. Because mutant alleles can only be descended from mutant alleles, the number of copies of the mutant allele (including descendants which have subsequently mutated) must be used for the population size in the coalescent equation. The number of copies of the mutant allele is assumed to increase

exponentially in a deterministic manner (which assumption is justified since the coalescent begins after the mutation has significant frequency). Because there cannot be selection for a mutation after fixation, the coalescent is only calculated back from fixation of an allele.

Preceding calculations, it is appropriate to clarify that there are three distinct numbers reflecting "population size": The actual population size is N, which is assumed constant; the number of copies of the mutant allele in the population (which is used in place of N in the coalescent equation); and the number of alleles in the coalescent (which is denoted by k). It also merits noting that the coalescent and fixation processes can be divided into two phases (Campbell 1999). The fixation process can be divided into the time from the first appearance of the mutation until the most recent common ancestor (MRCA) of the population and the time from the MRCA of the population to the present (i.e., fixation). The Coalescent process can be divided into the time (going backward) from the present to the fixation of the MRCA and the time from fixation of the MRCA to the MRCA. Since selection ends at fixation, only the time from fixation of the MRCA to the MRCA is calculated here. This is the same as the fixation time from the MRCA to the present (i.e., fixation).

The coalescent time

The differential form of the equation governing the coalescent

$$\frac{dk}{dt} = \frac{-k(k-1)}{2N} \tag{1}$$

where N is the population size and k is the number of alleles in the coalescent must be modified by replacing N with the number of copies of the mutant gene in the population. Assuming a constant (until fixation) Malthusian parameter s for the mutation, the frequency of the mutation will increase exponentially, hence the frequency at time t (going backward from fixation at $t = 0$; going backward in time requires the negative sign in the differential equation) is Ne^{-st} where N is the population size (which is assumed constant). The differential equation thereby becomes

$$\frac{dk}{dt} = \frac{-k(k-1)}{2Ne^{-st}}. \tag{2}$$

Solving this with the initial condition $k(0) = N$ yields the solution

$$\ln \frac{k}{k-1} = \frac{e^{st}-1}{2Ns} + \frac{1}{N} \tag{3}$$

which blows up at $k = 1$. However, the MRCA can also be interpreted as occurring one generation before $k = 2$ by which interpretation the MRCA occurred $\frac{\ln(\ln(2)2Ns-2s+1)}{s} + 1$ generations before the present. If a value other than $k = 2$ is used, the factor $\ln 2$ will be changed, for example $k = 1.59$ would replace $\ln 2$ with 1. Such changes are $O(\frac{1}{s})$, so allowing for freedom in the choice of k, the coalescent time is approximately

$$\frac{\ln 2Ns}{s} \tag{4}$$

for $Ns >> 1$.

At this time of the MRCA, $Ne^{-s\frac{\ln 2Ns}{s}} \doteq \frac{1}{2s}$, which is a large number, hence the assumption of deterministic population growth after the time of the MRCA is quite reasonable.

Consistency with the neutral case

In the absence of selection, it is also true that individuals can only be descended from descendants of the common ancestor. This smaller population size will result in a smaller coalescent time

than $2N$. In particular, the branching process model provides that a neutral allele (conditioned on non-extinction) increases linearly at one-half individual per generation (because the probability of extinction is approximately $1 - \frac{1}{2t}$ (Holte 1974) and the expected number of individuals is constant). Therefore, there are $N - \frac{t}{2}$ copies of a neutral mutation t generations before fixation, and the differential equation governing the coalescent time becomes

$$\frac{dk}{dt} = \frac{-k(k-1)}{2(N - t/2)}. \tag{5}$$

Solving this with the initial condition $k(0) = N$ and using $k = 2$ provides the time $t = N$ for the MRCA . Manifestly, $N < 2N$. However, this is not a contradiction. By employing the reduced size of the ancestral population, we are calculating the time between the MRCA and fixation as in the case of selection. The number $2N$ allows generations after fixation which share the same MRCA. Simulations of populations with 100 and 200 individuals (1000 trials for each size) showed that the number of generations between the present and fixation, and the number of generations between fixation and the most recent common ancestor are about equal (102 *versus* 97 and 195 *versus* 195). Hence N is the value which we should expect to get.

Fixation time

The fixation process is essentially the inverse of the coalescence process (in particular, both include the increase from MRCA to fixation). But although the generations after fixation of the most recent ancestor are not meaningful under selection (since selection ceases at fixation), the generations preceding the MRCA are meaningful in the context of selection. Hence the fixation time under selection is indeed the analogue of the neutral fixation time.

In order to use the above estimate $\frac{\ln 2Ns}{s}$ of the time from MRCA to fixation, it must be supplemented by the time from the occurrence of a beneficial mutation until the MRCA of the fixed population. This is the time until there are two lineages which do not go extinct (before fixation). If we condition on the lineage which never goes extinct (i.e., the skeleton; O'Connell 1993), under the Poisson progeny distribution (Karlin and Taylor 1975) the expected number of siblings for an individual in that lineage is $1 + s$. We may approximate the probability that the lineage of one of those individuals persists until fixation by the probability that a branching process does not go extinct, which is $2s$. Hence the expected time from the occurrence of a mutation until two descendants occur which will persist until fixation is

$$\sum_{i=0}^{\infty} (1 - (1+s)2s)^i = \frac{1}{2s(1+s)}. \tag{6}$$

The time from occurence until splitting into two immortal lines $\frac{1}{2s(1+s)}$ must be added to the time from the MRCA to fixation $\frac{\ln 2Ns}{s}$ to get the fixation time $\frac{\ln 2Ns}{s} + \frac{1}{2s(1+s)}$. However, the latter summand is the same order as the effect of choosing different values of k on determining the coalescent time, hence including it does not provide a more precise estimate. But it is worth noting that although in the neutral case the time from first appearance until the most recent common ancestor is approximately equal to the time from the MRCA until fixation, under strong selection the former is much less than the latter.

Shape of the coalescent and cumulative size

The structure (topology) of the coalescent is rather robust as long as all individuals are equivalent (different selection coefficients among members of the coalescent may impact the structure). Hence the present concern is at what rate does $k(t)$, the number of individuals in the coalescent at time t (the reduced branching process or reduced family tree; O'Connell 1995) increase. The rate

Table 1: Relative coalescent times and sizes

	$\frac{\frac{\ln 2Ns}{s}}{2N}$	(selection coalescent size)/(neutral coalescent size)					
		$N = 10^2$	10^3	10^4	10^5	10^6	10^7
$Ns = 10$	0.15	0.47	0.60	0.69	0.75	0.79	
$Ns = 100$	0.026		0.32	0.45	0.55	0.62	0.68

of growth of the coalescent is inversely proportional to the number of mutant genes in the population as reflected in equations (1) and (2), Hence the rate of change is accelerated most when the mutation is rare (and the coalescent is small), and least when the allele is near fixation (and $k(t)$ is near N). If one views the shape of the coalescent as a long stemmed rose in bloom (Campbell 2003), the effect of selection is to trim the stem but have little impact on the flower.

This nonuniform reduction of the rate of coalescence results in a great discrepancy between the reduction in the coalescent time (time between fixation and the MRCA of the fixed population) and the reduction of the cumulative size of the coalescent (total number of individuals in the coalescent from the MRCA until fixation). Results for strong selection ($Ns = 10$ and 100) for various population sizes ($N = 10^2, 10^3, 10^4, 10^5, 10^6, 10^7$) are presented in table 1.

The ratio of the fixation time for a mutation subject to strong selection to fixation time for a neutral mutation is $\frac{\ln 2Ns}{s}/2N = \frac{\ln 2Ns}{2Ns}$ which is a function of $2Ns$. For $Ns = 10$, this ratio is 0.15, entailing a 85% reduction in the coalescent time due to selection. For $Ns = 100$, this ratio is 0.026, entailing a 97% reduction in the coalescent time.

The cumulative size of the coalescent was obtained by solving (3) and (5) for k as a function of t, and then summing from $t = 0$ until the expected coalescent time. For the case of strong selection, $(1 - e^{\frac{1-e^{st}}{2Ns} - \frac{1}{N}})^{-1}$ is summed from 0 to $\frac{\ln 2Ns}{s}$. For the neutral case, $(1 - \frac{N-\frac{t}{2}}{N} e^{-\frac{1}{N}})^{-1}$ is summed from 0 to $2N$. (These solutions employed the appproximation $\ln(1 - \frac{1}{N}) = \frac{1}{N}$.) It merits mention that the cumulative coalescent sizes obtained by summation for the neutral case are all within 4% of the estimate $2N(\ln(2N) - 0.5)$ (Campbell 2003). The ratios displayed in Table 1 show a reduction in the cumulative coalescent size ranging from 68% to 21%, which is much less than the reduction in the coalescent time.

Logistic increase of favored allele

The analysis above has assumed constant viability of the selected allele until fixation. Logistic growth or relative viability models have produced the estimate for fixation time $\frac{2\ln(2N)}{s}$ (Campbell 1999; Durrett 2002). Qualitatively, logistic growth begins like exponential growth, but slows down at the end, hence will increase both the coalescent time and the cumulative size of the coalescent. Indeed, $\frac{2\ln(2N)}{s} > \frac{\ln(2Ns)}{s}$ (assuming $s < 1$). But $\frac{2\ln(2N)}{s}$ exceeds $2N$ for $s = 10^{-4}$ with $N = 10^5$, hence is not a good approximation for $Ns = 10$ in Table 1. For the values in Table 1 with $Ns = 100$, $\frac{2\ln(2N)}{s}$ ranges from 7.5% to 17% of $2N$, hence the approximation may be reasonable. There is a greater reduction in this coalescent time than in the cumulative coalescent size as calculated for the table, and the cumulative coalescent sizes in the table are less than under logistic increase; hence this supports that the cumulative coalescent size is reduced much less than the coalescent time under relative viability (logistic increase). (The differential equation governing the cumulative coalesent size under logistic growth is difficult to solve.)

Coalescent time for two alleles

The coalescent time for two alleles (rather than the entire population) was calculated numerically ($1 + \sum_{i=1}^{(\ln N)/s-1} \prod_{t=1}^{i} (1 - \frac{1}{Ne^{-ts}})$). ($\ln N/s$ is the number of generations to increase exponentially from 1 to N copies, and Ne^{-ts} is the number of selected alleles at time t.) The results for

$s = .1, .01, .001, .0001$ with $Ns = 10$ and 100 were approximately 20% and 4%, respectively, of the neutral time (N), hence entailed reductions of 80% and 96% (which are less than the reduction in the coalescent time for the entire population). The coalescent times for two individuals under selection can also be characterized as ranging from 67% to 80% of the coalescent time for the entire population under selection. This is much greater than 50% of the coalescent time for the entire population which occurs under neutrality. However, this does not reflect a star topology, the coalescent structure is robust. Rather it reflects the differential reduction in the rate of traversal for different portions of the coalescent.

Discussion

The primary result of this paper is that strong selection decreases genetic variation (measured as the number of alleles segregating in the population rather than heterozygosity) far less than it decreases the coalescent time. In the case of zero recombination (e.g., haploid selfing organisms such as bacteria; Dykhuizen 1990), genetic variation in the population ensues from mutations which occurred after the most recent common ancetor and in the coalescent, other variation is eliminated by the selective sweep (Barton 2000). Hence the number of mutations (mutations other than the selected one are assumed to be neutral, but mutations of the selected mutation share its selective advantage; the selected mutation loses its selective advantage once it is fixed) in the population will be proportional to the cumulative size of the coalescent, because that gives the amount of time (weighted by number of individuals) when mutations could occur. Although strong selection drastically reduces the coalescent time, it has little effect on the cumulative coalescent size, with reductions in Table 1 between 20% and 70%. Scrutinizing table 1 reveals that although Ns is the standard index of strength of selection and Ns governs the reduction in coalescent time, s is important for reduction in the size of the cumulative coalescent, with larger s entailing larger reduction. The cumulative coalescent size rather than the coalescent time is important for genetic variation, and since under strong selection the former is reduced far less than the latter, selection has a much smaller effect on genetic variation than might be assumed.

Another result relates to parsing the fixation time. The fixation time from occurrence of a mutation to its fixation can be parsed into the time from occurrence until the most recent common ancestor (MRCA) of the population where it is fixed plus the time from the MRCA of the population where it is fixed to fixation. Under selective neutrality the above parsing is essentially an equal division: $N + N = 2N$ provides the fixation time. Under strong selection, the fixation time is essentially equal to the time from the MRCA of the fixed population until fixation, the time from first occurrence of the mutation until the MRCA of the fixed population is much smaller. These times are of course expected values, the times will have large variances in practice.

Constant absolute rather than relative viability has been assumed for this model. However, this has limited significance. The change in the rate of coalescence is greatest when the mutation is rare, which is when the relative viability and absolute viability are approximately equal. When the mutation is frequent so that relative viabilities would be more appropriate, the frequency of the mutant is close to the population size, so the coalescent behaves approximately as the neutral case. The available estimate for the coalescent time under strong selection with relative viabilities does not produce reasonable values for the parameter values considered here.

Literature Cited

Barton, N. H. 2000. Genetic hitchhiking. Philos. Trans. R. Soc. Lond. B 355,1553-1562.

Barton, N. H. and A. M. Etheridge. 2004. The effect of selection on genealogies. Genetics 166,1115-1131.

Campbell, R. B. 1999. The coalescent time in the presence of background fertility selection. Theor. Popul. Biol. 55,260-269.

Campbell, R. B. 2003. A logistic branching process for population genetics. J. Theor. Biol. 225,195-203.

Donnely, P. and S. Tavaré. 1995. Coalescents and genealogical structure under neutrality. Annu. Rev. Genet. 29,401-421.

Durrett, R. 2002. Probability Models for DNA Sequence Evolution. Springer-Verlag. New York.

Dykhuizen, D. E. 1990. Experimental studies of natural selection in bacteria. Annu. Rev. Ecol. Syst. 21,373-398.

Ewens, W. J. 1979. Mathematical Population Genetics. Biomathematics vol. 9. Springer-Verlag. New York.

Griffiths, R. C. and S. Tavaré. 1994. Sampling theory for neutral alleles in a varying environment. Phil. Trans. R. Soc. Lond. B 344,403-410.

Holte, J. M. 1974. Extinction probability for a critical general branching process. Stoch. Proc. Appl. 2, 303-309.

Kaj, I. and S. M. Krone. 2003. The coalescent process in a population with stochastically varying size. J. Appl. Prob. 40,33-48.

Kaplan, N. L., T. Darden, and R. R. Hudson. 1988. The coalescent process in models with selection. Genetics 120,819-829.

Karlin, S. and H. M. Taylor. 1975. A First Course in StochasticProcesses (second edition). Academic Press. New York.

Kimura, M. and T. Ohta. 1969. The average number of generations until fixation of a mutant gene in a finite population. Genetics 61,763-771.

Kingman, J. F. C. 1982a. The coalescent. Stoch. Proc. Applicat. 13,235-248.

Kingman, J. F. C. 1982b. On the genealogy of large populations. J. Appl. Probability 19A,27-43.

Krone, S. M. and C. Neuhauser. 1997. Ancestral processes with selection. Theor. Popul. Biol. 51,210-237.

Neuhauser, C. and S. M. Krone 1997. The genealogy of samples in models with selection. Genetics 145, 519-534.

O'Connell, N. 1993. Yule process approximation for the skeleton of a branching process. J. Appl. Prob. 30, 725-729.

Polanski, A., A. Bobrowski, and M. Kimmel. 2003. A note on distributions of times to coalescence, under time-dependent population size. Theor. Popul. Biol. 63, 33-40.

Przeworski, M. 2003. Estimating the time since the fixation of a beneficial allele. Genetics 164, 1667-1676.

Sano, A., A. Shimizu, and M. Iizuka. 2004. Coalescent process with fluctuating population size and its effective size. Theor. Popul. Biol. 65, 39-48.

Slatkin, M. 1996. Gene genealogies within mutant allelic classes. Genetics 143, 579-587.

Slatkin, M. and R. R. Hudson. 1991. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics 129, 555-562.

van Herwaarden, O. A. and N. J. van der Wal. 2002. Extinction time and age of an allele in a large finite population. Theor. Popul. Biol. 61, 311-318.