STOCHASTIC MODELS IN POPULATION GENETICS

Shuhei Mano

DOCTOR OF
PHILOSOPHY

Department of Biosystems Science
School of Advanced Sciences
The Graduate University for Advanced Studies

2008

Abstract. Stochastic models have played important roles in population genetics. They have given understanding on evolutionary mechanism of maintaining genetic diversity within and between species. In this dissertation, the author will present several analytical results on stochastic models in population genetics, which have been obtained by the author and coworkers. The models cover various aspects, but with special reference to multi-locus models (Chapters 2, 3, 5) and the models with natural selection (Chapters 3, 4, 5). They are central issues under development in the current population genetics theory. With respect to random genetic drift for the one-locus problem, the state of steady decay was first obtained correctly by Wright (1931). Kimura (1955) obtained the complete expression of the transient probability density, which shows how the process leads to the state of steady decay. For two-locus problems, however, how the process eventually leads to the state of steady decay had not been studied. Two-locus problems are uniquely characterized by gamete frequencies. In Chapter 2, the conditional expectation of the transient gamete frequency, given that one of the two loci remains segregating, is obtained in terms of a neutral two-locus diffusion model. The sizes of natural populations change often. We are often interested in what population of constant size would have the same decrease in heterozygosity. This size is referred to as the effective size of the population. Wright (1938) pointed out that the effective size is approximately the harmonic mean of the individual sizes over the time period involved. For two-locus problems, Slatkin (1994) conjectured that, if a rapidly growing population is founded by a small size in which there is already linkage disequilibrium between a particular pair of loci, then closely linked loci will remain in significant linkage disequilibrium for a long time. However, no definite conclusions had been obtained, since there was no analytical framework for considering the effect. In Chapter 2, an asymptotic formula for the squared standard linkage deviation after a large number of generations is obtained in terms of a time-inhomogeneous stochastic model. According to the formula, in exponentially growing populations linkage disequilibrium will be asymptotically the same as that in a constant size population, the size of which is the current size. The evolutionary rate of a gene is defined as the rate of nucleotide substitutions. The rate is given by the product of the mutation rate and the fixation probability. Fixations of mutations also occur in genes that belong to a multigene family. It is possible that a mutation spreads over all member of a multigene family when they undergo concerted evolution, a phenomenon that the members evolve in a concerted manner by exchanging their DNA sequences. In Chapter 3, the rate of nucleotide substitutions in duplicated genes or a small multigene family, that are currently undergoing concerted evolution by gene conversion is investigated. A directional selection model, in which selection operates on the copy number of the mutant in a diploid, is investigated. The fixation probability is obtained in terms of a two-locus diffusion model. When no dominance exists among the selection coefficients, the formula can be extended to the n-locus model. According to the formula, the rate of molecular evolution is proportional to the size of the multigene family. It is known that GC-rich regions include many genes in mammalian genomes. A possible evolutionary force that might explain the pattern is biased gene conversion. Since biased mismatch DNA repair toward GC has been observed experimentally, gene conversion could favor particular alleles over others, or GC over AT base pairs. In fact, among multigene families undergoing concerted evolution, ribosomal operons, transfer RNAs, and histones are all GC-rich. In Chapter 3, a model of biased gene conversion is investigated. The fixation probability is obtained in terms of a neutral n-locus diffusion model. According to the formula, when the conversion rate is high, the acceleration of the rate of molecular evolution is proportional to square of the size of the multigene family. An ancestral genealogy of a sample of genes plays an important role in a probabilistic description of the sample. The size process, which is the the number of ancestors backward in time of a sample, is referred to as the ancestral process. The ancestral selection graph introduced by Krone and Neuhauser (1997) is an analogue of the coalescent genealogy. Few properties were known about the ancestral process, which is the total number of ancestral particles in a cross section of an ancestral selection graph backward in time of a sample. In Chapter 4, properties of the ancestral process are investigated. The probability distribution is obtained by using a moment dual relationship between the ancestral process and a diffusion model investigated by Kimura (1955). Bounds for the probability that the ancestral process is at the state one are obtained by an elementary martingale argument, which is an extension of the bounds obtained by Kingman (1982) for the neutral process. It is shown that the process of fixation of the allele in the diffusion model corresponds to convergence of the ancestral process to its stationary measure. Developing statistical methods to detect adaptive evolution with DNA sequence data has been an important issue. The methods using within species polymorphism data can be loosely classified into two categories: site frequency methods and haplotype frequency methods. The site frequency methods require only frequencies of variants at polymorphic nucleotide sites. In contrast, the haplotype frequency methods require additional information on the linkage phases among variant sites. Recently, the author and coworkers (2007) found that the Watterson's homozygosity test (1978) is usually robust against intra-haplotype recombination and the most powerful test during the sweep phase. However, the test is based on a summary statistic and gives few insights how a selection operates. In Chapter 5, a new likelihood based test to detect recent sweeps with utilizing haplotype frequency data is presented. The test provides maximum likelihood estimates of the position and intensity of the target of a selection.

Acknowledgments

I am grateful to the chairman of my supervisory committee, Dr. Masami Hasegawa, and members of my supervisory committee, Drs. Hideki Innan, Toshiyuki Takano-Shimizu, and Takashi Gojobori, for their valuable advices.

I am also grateful to Dr. Hideki Innan for a collaborative work and for his help through accomplishment of this dissertation.

Finally, I express my gratitude to Dr. Takashi Gojobori, who introduced me to the research area, for his continuous encouragement.

September 30, 2008

Contents

CHAPTER 1

Introduction

Stochastic models have played important roles in population genetics. They have given theoretical understanding on evolutionary mechanism of maintaining genetic diversity within and between species. Following in a line of Fisher (1930) and Wright (1945), in 1950-1980s Kimura and his coworkers had given foundations of theories of molecular evolution by developing stochastic models based on the diffusion process. By applying their theoretical predictions to emerging molecular data at that time, various important aspects of molecular evolution have been revealed. The most significant prediction is probably the neutral hypothesis of molecular evolution, which was advocated by Kimura (1968). In 1982-1983, a stochastic model, which is now called the coalescent model, was introduced (Kingman, 1982b; Tajima, 1983; Hudson, 1983b). The coalescent process is a stochastic process of ancestors of a sample of genes, which are taken from a population evolving under the diffusion model. The coalescent model has given a useful framework of statistical analysis of a sample taken from a population.

In this dissertation, the author will present several analytical results on stochastic models in population genetics, which have been obtained by the author and coworkers. The models cover various aspects, but with special reference to multi-locus models (Chapters 2, 3, 5) and the models with natural selection (Chapters 3, 4, 5). They are central issues under development in the current population genetics theory.

With respect to random genetic drift for the one-locus problem, the state of steady decay was first obtained correctly by Wright (1931). By calculating moments of the distribution, Kimura (1955a) obtained the complete expression of the transient probability density for the unfixed class, which shows how the process leads to the state of steady decay. For two-locus problems, however, how the process eventually leads to the state of steady decay had not been studied, with the exception of several functions (Ohta and Kimura, 1969a). Two-locus problems are uniquely characterized by gamete frequencies. In Chapter 2, an analytic expression of conditional expectation of the transient gamete frequency, given that one of the two loci remains segregating, is obtained in terms of a two-locus diffusion model. Using this expression, a model where linkage disequilibrium is

introduced by a single mutation is discussed. The behavior of the conditional expectation of gamete frequency is significantly different from the monotonic decrease observed in the deterministic model without random genetic drift. The results were published in Mano (2005).

The sizes of natural populations change often. We are often interested in what population of constant size would have the same decrease in heterozygosity. This size is referred to as the effective size of the population. Wright (1938) pointed out that the effective size is approximately the harmonic mean of the individual sizes over the time period involved. This means that a single period of small population size, called a bottleneck, can result in a significant decrease in heterozygosity (Nei et al., 1975). For two-locus problems, Slatkin (1994a) conjectured that, if a rapidly growing population is founded by a small size in which there is already linkage disequilibrium between a particular pair of loci, then closely linked loci will remain in significant linkage disequilibrium for a long time. The fate of linkage disequilibrium which already exists in the founder population has practical importance for designing association analyses for mapping complex traits genes (Lander and Botstein, 1986; Laan and Pääbo, 1997). Nevertheless, no definite conclusions had been obtained, since there was no analytical framework for considering effects of change of population sizes on linkage disequilibrium. In Chapter 2, evolution of linkage disequilibrium of the founders in exponentially growing populations is investigated in terms of a time-inhomogeneous stochastic model. As a measure of linkage disequilibrium, the squared standard linkage deviation is considered. By a perturbative series expansion in a growth parameter, an asymptotic formula for the squared standard linkage deviation after a large number of generations is obtained. According to the formula, in exponentially growing populations, linkage disequilibrium will be asymptotically the same as that in a constant size population, the effective size of which is the current size. The results were published in Mano (2007).

The evolutionary rate of a gene is defined as the rate of nucleotide substitutions (Zuckerkandl and Pauling, 1965; Jukes and Canter, 1969). The rate is given by the product of the mutation rate and the fixation probability. Fixations of mutations also occur in genes that belong to a multigene family. It is possible that a mutation spreads over all member genes of a multigene family when they undergo concerted evolution, a phenomenon that the members evolve in a concerted manner by exchanging their DNA sequences (Ohta, 1980; Dover, 1982). In Chapter 3, the rate of nucleotide substitutions in duplicated genes or a small multigene family, that are currently undergoing concerted evolution by gene

conversion is investigated. Gene conversion between copy members should be the major mechanism to cause concerted evolution of small multigene families (Ohta, 1983a). A directional selection model, in which selection operates on the copy number of the mutant in a diploid, is investigated. An analytic expression of the fixation probability is obtained in terms of a two-locus diffusion model. When no dominance exists among the selection coefficients, the formula for the fixation probability can be extended to the n-locus model. Interestingly, the formula is identical to the formula for the fixation probability of a mutant with genic selection in a subdivided population (Maruyama, 1972). According to the formula, selection will operate more efficiently in a large multigene family; the rate of molecular evolution is roughly proportional to the size of the multigene family. The results were published in Mano and Innan (2008).

It is known that GC-rich regions include many genes in mammalian genomes (Duret et al., 1995). A possible evolutionary force that might explain the pattern is biased gene conversion. Since biased mismatch DNA repair toward GC has been observed experimentally (Brown and Jiricny, 1987), gene conversion could favor particular alleles over others, or GC over AT base pairs. If biased gene conversion were major determinant of GC content evolution, one would expect sequences undergoing frequent gene conversion to become GC-rich. In fact, among multigene families undergoing concerted evolution in mammals, ribosomal operons, transfer RNAs, and histones are all GC-rich, consistent with the prediction (Galtier, et al. 2001). In Chapter 3, a model of biased gene conversion is investigated. An analytic expression of the fixation probability is obtained in terms of an n-locus diffusion model. According to the formula, the bias in gene conversion will have significant effect upon a large multigene family; when the conversion rate is large, the acceleration of the rate of molecular evolution is proportional to square of the size of the gene family.

An ancestral genealogy of a sample of genes plays an important role in a probabilistic description of the sample. Let $a_n(t)$ be the number of ancestors at time t backward of a sample of n neutral genes. The size process is referred to as the ancestral process. The distribution of $a_n(t)$ is known (Griffiths (1979), Tavaré (1984)). The ancestral selection graph introduced by Krone and Neuhauser (1997) is an analogue of the coalescent genealogy. The elements are referred to as particles. Let $b_n(t)$ be the number of edges, or ancestral particles, in a cross section of an ancestral selection graph at time t backward of a sample of n genes. In the case of no mutation, the real genealogy of a sample is the same as in the neutral process (Theorem 3.12 in Krone and Neuhauser (1997)). In

contrast, few properties of the ancestral process $\{b_n(t); t \geq 0\}$, which is the size process of the total number of the real and virtual particles, were known. In Chapter 4, properties of the ancestral process are investigated. An explicit form of the probability distribution is obtained, by using a dual relationship between the ancestral process and a diffusion model investigated by Kimura (1955c) in a context by Tavaré (1984). The ancestral process converges to the stationary measure, which is the truncated Poisson distribution. In contrast to the neutral process, the final rates of convergence are given by the largest eigenvalue for all the states. Bounds for the probability that the ancestral process is at the state one are obtained by an elementary martingale argument, which is an extension of the bounds obtained by Kingman (1982a) for the neutral process. By killing the modified process, the formal form of the joint probability generating function of the ancestral process and the number of branching events is obtained. It is shown that the process of fixation of the allele in the diffusion model corresponds to convergence of the ancestral process to its stationary measure. Especially, the density of time to fixation of a single mutant conditional on fixation is given by the probability of the whole population being descended from a single real ancestral particle, regardless of the allelic type. The results were presented in Mano (2008).

Developing statistical methods to detect adaptive evolution with DNA sequence data has been an important issue. The methods using within species polymorphism data can be loosely classified into two categories: site frequency and haplotype frequency methods. The site frequency methods require only frequencies of variants at polymorphic nucleotide sites. Linkage phase of these variants is not used. The methods are based on the completely linked infinite site model and utilize the simple summary statistics of site frequency spectrum (e.g., Tajima (1989a); Fu and Li (1993); Fay and Wu (2000)). The haplotype frequency methods require additional information on the linkage phase among variant sites. A haplotype is scored as an allele and conditional haplotype frequency spectrum are used for detection. One sub-category of the method is based on the infinite allele model and utilize allele frequency spectrum conditional on the number of different haplotypes (Ewens, 1973b; Watterson, 1978; Slatkin, 1994b). The other sub-category of the methods is based on the infinite sites model and utilize allele frequency spectrum conditional on the number of segregating sites (Depaulis and Veuille, 1998; Innan et al., 2005). Recently, the author and coworkers assessed the power and robustness of these haplotype and site frequency methods to detect positive selection by extensive simulations (Zeng et al., 2007). In their study, intra-haplotype recombination were incorporated. They found

that although the haplotype frequency methods conditional on the number of haplotypes were constructed based on the infinite allele model without recombination, these tests are insensitive to intra-haplotype recombination. It means that the number of haplotypes has information of both of mutation and recombination. In addition, they found that the Watterson's homozygosity test (Watterson, 1978) is usually the most powerful test during the sweep phase, especially when the local recombination rate is high. However, since the Watterson's homozygosity test is based on a summary statistic, it gives few insights how the selection operates. In contrast, likelihood based tests which utilize the site frequency spectrum at unlinked segregating sites (e.g., Kim and Stephan (2002); Nielsen et al. (2006)) can provide maximum likelihood estimates of the position of the target of selection and the selection intensity. In Chapter 5, a new likelihood based test to detect a recent sweep which utilizes haplotype frequency data is presented. The likelihood for the model at the end of the selective sweep, a sampling formula, was presented by the author (Mano, 2006).

CHAPTER 2

Linkage Disequilibrium

2.1. Introduction

With respect to random genetic drift for the one-locus problem, the state of steady decay was first obtained correctly by Wright (1931). However, in this study it was assumed that the state of steady decay had already been attained. By calculating moments of the distribution, Kimura (1955a) obtained the complete expression of the transient probability density for the unfixed class, which shows how the process leads to the state of steady decay. It was found that after $2N$ generations the distribution becomes almost flat, where N is the effective population size.

Since each mutant ultimately becomes either fixed or lost, the stationary state will be attained only if evolutionary pressures, such as mutation, operate. For two-locus problems, the stationary state has been discussed in terms of the diffusion process (Ohta and Kimura, 1969b; Griffiths, 1981; Ethier and Nagylaki, 1989; Ethier and Griffiths, 1990) and the genealogical process (Hudson, 1983a; Golding, 1984; Hudson, 1985). In contrast, situations without evolutionary pressures, how the process eventually leads to the state of steady decay has not been studied, with the exception of several functions which vanish at the absorbing boundaries (Hill and Robertson, 1968; Ohta and Kimura, 1969a; Litter, 1973). Despite the fact that two-locus problems are uniquely characterized by gamete frequencies, the transient behavior of them has not been examined.

In this chapter, an analytic expression of conditional expectation of the transient gamete frequency, given that one of the two loci remains segregating, in terms of the diffusion process is presented. The expression was obtained by the author (Mano, 2005). This expression shows how the process leads to the state of steady decay. Using this expression, a model where linkage disequilibrium is introduced by a single mutation is discussed.

The sizes of natural populations change often. These changes play important roles in population genetics. We are often interested in what population of constant size would have the same decrease in heterozygosity. This size is referred to as the effective size of

the population. Wright (1938) pointed out that the effective size is approximately the harmonic mean of the individual effective sizes over the time period involved. This means that a single period of small population size, called a bottleneck, can result in a significant decrease in heterozygosity (Nei et al., 1975).

Recently, to infer change of population sizes from polymorphism data, effects of change of population sizes on various statistics, such as nucleotide site differences in pairwise comparisons of DNA sequences (Li, 1977; Tajima, 1989b; Slatkin and Hudson, 1991; Rogers and Harpending, 1992), and microsatellite repeat variability (Kimmel et al., 1998; Reich and Goldstein, 1998; Thomson et al., 2000) were studied. By simulations, Slatkin (1994a) showed that in a rapidly growing population there is little chance of detecting linkage disequilibrium between completely linked loci. However, in his simulations all of the polymorphisms were assumed to have arisen by mutations after the population was founded. He did not consider the evolution of linkage disequilibrium which already existed in the founder population. It was conjectured that, if a population is founded by a small size in which there is already linkage disequilibrium between a particular pair of loci, then very closely linked loci will remain in significant linkage disequilibrium for a long time. In addition, the fate of linkage disequilibrium which already exists in the founder population has practical importance for designing association mapping methods for complex traits genes (Lander and Botstein, 1986; Laan and Pääbo, 1997). Several studies based on simulations were conducted so far (Terwilliger et al., 1998; Kruglyak, 1999). Nevertheless, no definite conclusions have been obtained, since there is no analytical framework for considering the effects of change of population sizes on linkage disequilibrium.

In this chapter, evolution of linkage disequilibrium which already exists in the founders of exponentially growing populations is studied, which was presented by the author (Mano, 2007). The properties of the squared standard linkage deviation, which is defined by the ratio of the moments, are considered, analytically, numerically and by simulations. By using the diffusion approximation of the Wright-Fisher model, Ohta and Kimura (1969a,b) studied evolution of the squared standard linkage deviation in constant size populations. Here, the squared standard linkage deviation in exponentially growing populations is studied by using a time-inhomogeneous diffusion model, which is an approximation of the time-inhomogeneous Wright-Fisher model, where the population size grows exponentially in a deterministic way.

2.2. A two-locus diffusion model

Consider a random mating population with an effective population size of N. We will measure time t in units of $2N$ generations. Let A_1 and A_2 be a pair of alleles with initial frequencies are p and $1-p$, respectively, and the allele frequencies at time t are x and $1-x$, respectively. A diffusion time scaling is to let $2N \to \infty$. The Wright-Fisher model converges to a diffusion process. Kimura (1955a) obtained an analytic expression of the transient probability density for the unfixed class. Let $\phi(p, x; t)$ be the probability density. The probability that the locus remains segregating was also given;

$$\begin{aligned} \mathbb{P}[x \in (0,1)] &= \int_0^1 \phi(p,x;t)dx = 1 - \lim_{n\to\infty} E[x^n] - \lim_{n\to\infty} E[(1-x)^n] \\ &= \sum_{m=0}^{\infty} \left\{P_{2m}(1-2p) - P_{2m+2}(1-2p)\right\} e^{-\frac{(2m+1)(2m+2)}{2}t}, \end{aligned} \tag{2.1}$$

where $P_m(z)$ represents the Legendre polynomial. In general, since we cannot observe polymorphisms that have been lost, we have interest in the conditional expectation of the frequencies given that the locus remains segregating. By using the expression of the transient fixation probability given by Kimura (1955a), we have the conditional expectation of the allele frequency for the unfixed class

$$\mathbb{E}[x|x \in (0,1)] = \frac{\mathbb{E}[x, x \in (0,1)]}{\mathbb{P}[x \in (0,1)]}, \tag{2.2}$$

where

$$\begin{aligned} \mathbb{E}[x, x \in (0,1)] &= \int_0^1 x\phi(p,x;t)dx = \mathbb{E}[x] - f(1;t) \\ &= \sum_{m=1}^{\infty} \frac{(-1)^m}{2} \left\{P_{m+1}(1-2p) - P_{m-1}(1-2p)\right\} e^{-\frac{m(m+1)}{2}t}, \end{aligned}$$

where $f(1;t)$ represents the transient fixation probability of the allele A_1. The asymptotic value of the conditional expectation of the allele frequency is

$$\mathbb{E}[x|x \in (0,1)] \to \frac{1}{2}, \qquad t \to \infty, \tag{2.3}$$

which agrees with the fact that the conditional distribution becomes to uniform asymptotically.

Let us assume two loci A and B in which pair of alleles A_1, A_2 and B_1, B_2 are segregating, and let the initial frequencies of gametes A_1B_1, A_1B_2, A_2B_1, and A_2B_2 be respectively g_1, g_2, g_3, and $1-(g_1+g_2+g_3)$, and let the frequencies of them at time t be respectively x_1, x_2, x_3, and $1-(x_1+x_2+x_3)$. Let the initial frequencies of alleles B_1 and B_2 be respectively q and $1-q$, and the frequencies of them at time t be respectively y and $1-y$. Let

$D = g_1(1-g_1-g_2-g_3)-g_2g_3$ be the initial value of the linkage disequilibrium coefficient and $z = x_1(1-x_1-x_2-x_3)-x_2x_3$ be the value of the linkage disequilibrium coefficient at time t. We have

$$x_1 = xy+z, \qquad x_2 = x(1-y)-z, \qquad x_3 = (1-x)y-z. \tag{2.4}$$

Let r be the recombination rate between the loci. We will not discuss where $r=0$, since the problem reduces to the multi-allelic one-locus problem which has previously been discussed by Kimura (1955b). For the deterministic model without random genetic drift, we have $x=p$, $y=q$, and $z = De^{-2Nrt}$.

A diffusion time scaling is to measure time in units of $2N$ generations and let $2N\to\infty$, while $\rho = 4Nr$ is held constant. The Wright-Fisher model converges to a diffusion process. The probability density for the gamete frequencies $\phi(g_1,g_2,g_3;x_1,x_2,x_3;t)$ satisfies the following Kolmogorov backward equation (Ohta and Kimura, 1969a),

$$\frac{\partial\phi}{\partial t} = \sum_{i,j=1}^{3}\frac{g_i(\delta_{ij}-g_j)}{2}\frac{\partial^2\phi}{\partial g_i\partial g_j} - \frac{\rho D}{2}\left(\frac{\partial\phi}{\partial g_1}-\frac{\partial\phi}{\partial g_2}-\frac{\partial\phi}{\partial g_3}\right), \tag{2.5}$$

where δ_{ij} represents the Kronecker's delta. The forward equation of the process was firstly obtained by Hill and Robertson (1966). Although the probability density is unknown, Ohta and Kimura (1969a) obtained expectations of functions

$$x(1-x)y(1-y), \qquad (1-2x)(1-2y)z, \qquad z^2, \tag{2.6}$$

which were discussed by Hill and Robertson (1968). The process is defined in a simplex

$$K : 0 \le x_1 \le x_1+x_2 \le x_1+x_2+x_3 \le 1. \tag{2.7}$$

When we define a map Φ by $\Phi(x_1,x_2,x_3) = (x,y,z)$ and letting $H = \Phi(K)$, Φ is a C^∞-diffeomorphism of K onto H. The upper part of ∂H is depicted in Figure 2.1. On the peripheral edges, which is the periphery of the square $0\le x\le 1, 0\le y\le 1$, either of the two loci is not segregating. At the points $(1,1,0),(1,0,0),(0,1,0)$, and $(0,0,0)$, either of the gametes A_1B_1, A_1B_2, A_2B_1, and A_2B_2 fixes respectively. The generator of the diffusion process $\{x(t),y(t),z(t);t\ge 0\}$, which is defined by $(x(t),y(t),z(t)) = \Phi(x_1(t),x_2(t),x_3(t))$ in H, is (Ohta and Kimura, 1969a)

$$\begin{aligned} L &= \frac{x(1-x)}{2}\frac{\partial^2}{\partial x^2}+\frac{y(1-y)}{2}\frac{\partial^2}{\partial y^2}+z\frac{\partial^2}{\partial x\partial y}+z(1-2x)\frac{\partial^2}{\partial x\partial z}+z(1-2y)\frac{\partial^2}{\partial y\partial z} \\ &\quad -z\left(1+\frac{\rho}{2}\right)\frac{\partial}{\partial z}+\frac{1}{2}\left\{xy(1-x)(1-y)+z(1-2x)(1-2y)-z^2\right\}\frac{\partial^2}{\partial z^2}. \end{aligned} \tag{2.8}$$

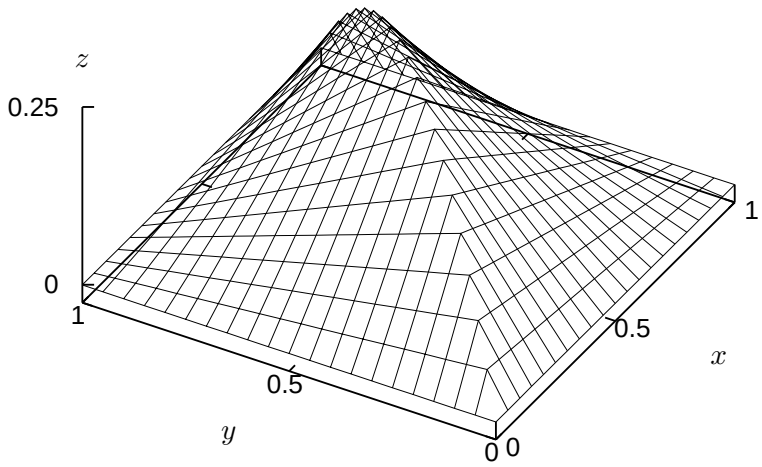

FIGURE 2.1. The upper surface of the boundary of the region in which the diffusion process $\{x(t), y(t), z(t); t \geq 0\}$ is defined.

The expectation of the linkage disequilibrium coefficient is (Hill and Robertson, 1968)

$$\mathbb{E}[z] = De^{-(1+\frac{\rho}{2})t}, \tag{2.9}$$

and the squared standard linkage deviation asymptotically tends to (Ohta and Kimura, 1969a)

$$\sigma_d^2 := \frac{\mathbb{E}[z^2]}{\mathbb{E}[x(1-x)y(1-y)]} \to \frac{1}{\rho} + O(\rho^{-2}), \qquad t \to \infty, \tag{2.10}$$

when ρ is large.

Let us discuss expectation of the gamete frequencies. In the same manner as for the functions (2.6) and the linkage disequilibrium measures, we obtain the expectation of the gamete frequency

$$\mathbb{E}[x_1] = g_1 + \frac{\rho D}{2+\rho}\left\{e^{-(1+\frac{\rho}{2})t} - 1\right\}. \tag{2.11}$$

However, in contrast to the functions (2.6) and the linkage disequilibrium measures, the gamete frequencies do not vanish at the peripheral edges. The expectation takes over not only the inside of the region, but also the boundaries ∂H. Thus, the expectation of the gamete frequency can be rewritten formally as

$$\mathbb{E}[X_1] = \int\int\int_{H-\partial H} x_1\phi dxdydz + \int_0^1 x_1\phi_{x=1}dy + \int_0^1 x_1\phi_{y=1}dx + f(1,0,0;t), \tag{2.12}$$

where $\phi_{x=1}$ and $\phi_{y=1}$ represent the probability density for the open intervals $x = 1, 0 < y < 1, z = 0$ and $0 < x < 1, y = 1, z = 0$, respectively. $f(1, 0, 0; t)$ represents the transient fixation probability of the gamete A_1B_1 at time t. Here, it is implicitly assumed that there are no probability at ∂H other than the peripheral edges, in which either of the four possible gametes are lost. We have no rigorous justification for the assumption, however, in biological point of view, the assumption seems to be natural; Because of recombination there are no possibility that a population stays at ∂H other than the peripheral edges.

2.3. Conditional expectation of gamete frequency

Suppose linkage disequilibrium is introduced by a single mutation, as considered by Nei and Li (1980) regarding the association between electromorphs and inversion chromosomes in *Drosophila*. We assume the locus A is not segregating and the wild type allele A_2, and the locus B, in which a pair of alleles B_1 and B_2 (electromorphs) are segregating with the allele frequencies q and $1-q$, respectively. Then, the mutation introduces the mutant allele (inversion chromosome) A_1 to the locus A of one of the allele B_1 bearing chromosomes. In this setting, the polymorphism at the locus A is critical since the allele A_1 is prone to be lost by random genetic drift. The locus B may be regarded as a marker polymorphism to detect the mutant.

Motivated by the example introduced above, we will consider the conditional expectation given that the locus A remains segregating. It might seem that this condition is similar to that described by Kaplan and Weir (1992). They discussed conditional expectation of a linkage disequilibrium measure, which was defined by Nei and Li (1980), given that polymorphism is observed at the locus B. They assumed that the allele frequency of A_1 is constant, and the locus B follows the infinite allele model. Moreover, they considered the stationary state. Thus, their model differs from that described here considerably, and the condition that the locus A remains segregating is critical for our discussion. Note that this condition nearly equates to a condition that both of the two loci remain segregating, since the probability that a fixation occurs at the locus A earlier than the locus B is given by (Karlin and McGregor, 1968) $q(1-q)/\{q(1-q)+p(1-p)\}$, which is almost unity unless q is very small.

By expression (2.12), we have

$$\begin{aligned}\mathbb{E}[x_1, x \in (0,1)] &= \iiint_{\mathcal{D}} x_1 \phi dx dy dz + \int_0^1 x_1 \phi_{y=1} dx \\ &= \mathbb{E}[x_1] - f(1,0,0;t) - \int_0^1 x_1 \phi_{x=1} dy \\ &= \mathbb{E}[x_1] - \lim_{n\to\infty} \mathbb{E}[x^n x_1] = \mathbb{E}[x_1] - \lim_{n\to\infty} \mathbb{E}[x^n y]. \end{aligned} \tag{2.13}$$

The expressions for the other gamete frequencies can be obtained in the same manner. To calculate the limit of the expectation $\lim_{n\to\infty} \mathbb{E}[x^n y]$, we will consider some moments. Let

$$\mu_{l,m,n} = \mathbb{E}[x^l y^m z^n], \qquad l, m, n = 0, 1, ... \tag{2.14}$$

Making use of the Itô formula with the generator (2.5) (See, Appendix), we have a differential equation for the moments

$$\begin{aligned}\frac{d\mu_{l,m,n}}{dt} &= -\frac{l(l-1)+m(m-1)+n(n-1)+n\{4(l+m)+2+\rho\}}{2}\mu_{l,m,n} \\ &+\frac{l(l-1+2n)}{2}\mu_{l-1,m,n} + \frac{m(m-1+2n)}{2}\mu_{l,m-1,n} + lm\mu_{l-1,m-1,n+1} \\ &+\frac{n(n-1)}{2}\{\mu_{l,m,n-1} + \mu_{l+1,m+1,n-2} - 2(\mu_{l+1,m,n-1} + \mu_{l,m+1,n-1}) \\ &-\mu_{l+1,m+2,n-2} - \mu_{l+2,m+1,n-2} + 4\mu_{l+1,m+1,n-1} + \mu_{l+2,m+2,n-2}\}. \end{aligned} \tag{2.15}$$

It is worthwhile to note that $\mathbb{E}[x^l y^m x_1^n]$ satisfies a recurrence relation which is the same as the recurrence relation for the two-locus sampling distribution (Golding, 1984; Ethier and Griffiths, 1990), which has a genealogical interpretation in terms of the two-locus ancestral recombination graph (Griffiths, 1991). Namely, for $\xi_{l,m,n} = \mathbb{E}[x^l y^m x_1^n] = \mathbb{E}[p^a q^b g_1^c]$,

$$\begin{aligned}\frac{d\xi_{l,m,n}}{dt} &= -\frac{(l+m+n)(l+m+n-1)+n\rho}{2}\xi_{l,m,n} + \frac{n\rho}{2}\xi_{l+1,m+1,n-1} \\ &+\frac{l(l-1+2n)}{2}\xi_{l-1,m,n} + \frac{m(m-1+2n)}{2}\xi_{l,m-1,n} + lm\xi_{l-1,m-1,n+1} \\ &+\frac{n(n-1)}{2}\xi_{l,m,n-1}, \end{aligned} \tag{2.16}$$

where $\{a(t), b(t), c(t); t \geq 0\}$ is a Markov process of the number of edges ancestral to a sample with $a(0) = l, b(0) = m, c(0) = n$. $a(t)$ is in the number of edges which are ancestral to the sample in the locus A only, $b(t)$ is the number of edges which are ancestral to the sample in the locus B only, and $c(t)$ is the number of edges which are ancestral to the sample in both of the loci.

The moments $\mu_{n,0,1}$ satisfy a system of differential equations

$$\frac{d\mu_{n-1,0,1}}{dt} = -\left\{\frac{n(n+1)+\rho}{2}\right\}\mu_{n-1,0,1} + \frac{n(n-1)}{2}\mu_{n-2,0,1}, \qquad n = 2, 3, ... \tag{2.17}$$

with the initial condition $\mu_{n-1,0,1}(0) = p^{n-1}D, n = 1, 2, ...$ It is straightforward to show that the solution has a form

$$\mu_{n-1,0,1}(t) = \sum_{m=1}^{n} C_{n-1}^{(m)}(p) D e^{-\frac{m(m+1)+\rho}{2}t}, \qquad n = 1, 2, ... \tag{2.18}$$

with

$$\begin{aligned} C_{n-1}^{(m)}(p) &= \frac{n(n-1)}{(n+m+1)(n-m)} C_{n-2}^{(m)}(p) = \cdots \\ &= \frac{n!(n-1)!(2m+1)!}{(n+m+1)!(n-m)!m!(m-1)!} C_{m-1}^{(m)}(p). \end{aligned} \tag{2.19}$$

The explicit form of $C_{m-1}^{(m)}(p)$ is given by the following lemma.

LEMMA 2.3.1.

$$C_{m-1}^{(m)}(p) = \frac{m!(m-1)!}{(2m)!} 2(-1)^{m+1} T_{m-1}^{1}(1-2p), \qquad m = 1, 2, ..., \tag{2.20}$$

where $T_m^1(z)$ represents the Gegenbauer polynomial, which is also represented as $C_m^{\frac{3}{2}}(z)$.

PROOF. The initial condition is

$$p^{n-1} = \sum_{m=1}^{n} \frac{n!(n-1)!(2m+1)!}{(n+m+1)!(n-m)!m!(m-1)!} C_{m-1}^{(m)}(p), \qquad n = 1, 2, ... \tag{2.21}$$

Since the Gegenbauer polynomial $T_m^1(z)$ is an orthogonal polynomial on the interval $[-1, 1]$, p^n should be represented in terms of the Gegenbauer polynomials whose degrees are up to $n-1$, it is possible to set that

$$C_{m-1}^{(m)}(p) = C_m T_{m-1}^{1}(r), \qquad r = 1 - 2p. \tag{2.22}$$

By multiplying $(1-r^2)T_{m-1}^1(r)$ on both sides of (2.21) and using the orthogonal property

$$\int_{-1}^{1} (1-z^2) T_{k-1}^{1}(z) T_{l-1}^{1}(z) dz = \delta_{kl} \frac{2l(l+1)}{2l+1}, \qquad k, l = 1, 2, ... \tag{2.23}$$

we have

$$\begin{aligned} C_m &= \frac{(-1)^{m+1}(n+m+1)!(n-m)!\{(m-1)!\}^2}{2^{n+1}n!(n-1)!(2m-1)!m(m+1)} \int_{-1}^{1} (1-r)(1+r)^n T_{m-1}^{1}(r) dr \\ &= \frac{\{(m-1)!\}^2}{(2m-1)!}(-1)^{m+1}, \end{aligned} \tag{2.24}$$

where an integral transform by the Gegenbauer polynomial for $n = 0, 1, ...; m = 1, 2, ...$ (Erdélyi, 1954)

$$\int_{-1}^{1} (1-z)(1+z)^n T_{m-1}^{1}(z) dz = \frac{2^{n+1}\{(n-1)!\}^2 nm(m+1)}{(n+m+1)!(n-m)!} \tag{2.25}$$

is employed. □

The moments $\mu_{n,1,0}$ satisfy a system of differential equations

$$\frac{d\mu_{n,1,0}}{dt} = -\frac{n(n-1)}{2}(\mu_{n,1,0} - \mu_{n-1,1,0}) + n\mu_{n-1,0,1}, \qquad n = 1, 2, ... \tag{2.26}$$

and the differential equation has the solution of the form for $n = 1, 2, ...$

$$\mu_{n,1,0}(t) = pq + \frac{D}{1+\frac{\rho}{2}} + \sum_{m=1}^{n-1} E_n^{(m)}(p,q,D)e^{-\frac{m(m+1)}{2}t} + \sum_{m=1}^{n} F_n^{(m)}(p)De^{-\frac{\rho+m(m+1)}{2}t}, \tag{2.27}$$

where

$$E_n^{(m)}(p,q,D) = \frac{n!(n-1)!(2m+1)!}{(n+m)!(n-m-1)!(m+1)!m!}E_{m+1}^{(m)}(p,q,D) \tag{2.28}$$

and

$$\{(n+m)(n-m-1) - \rho\}\, F_n^{(m)}(p) = n(n-1)F_{n-1}^{(m)}(p) + 2nC_{n-1}^{(m)}(p), \tag{2.29}$$

with the initial condition

$$p^n q = pq + \frac{D}{1+\frac{\rho}{2}} + \sum_{m=1}^{n-1} E_n^{(m)}(p,q,D) + \sum_{m=1}^{n} F_n^{(m)}(p)D, \qquad n = 1, 2, ... \tag{2.30}$$

The recurrence relation (2.29) can be expressed by using a matrix $\boldsymbol{Af} = \boldsymbol{c}$, where $f_k = F_k^{(m)},\ c_k = 2kC_{k-1}^{(m)},\ k = m, m+1, ..., n$. The determinant of the matrix A is

$$\det \boldsymbol{A} = \prod_{k=m}^{n} \{k(k-1) - m(m+1) - \rho\}, \tag{2.31}$$

which has zeros at $\rho = 2 + 2l,\ l = 1, 2, 3, ...$ These zeros are due to degeneracy of the eigenvalues. Since we are not interested in the specific points of ρ, we will discuss the case that the inverse matrix exists in the following, although the calculation with these zeros is straightforward. By applying the inverse matrix, we obtain

$$\begin{aligned} F_n^{(m)}(p) &= \sum_{k=1}^{n-m+1} \frac{2n!(n-1)!}{\{(n-k)!\}^2} \frac{\Gamma\left(n-k+\frac{1}{2}+\rho_m\right)\Gamma\left(n-k+\frac{1}{2}-\rho_m\right)}{\Gamma\left(n+\frac{1}{2}+\rho_m\right)\Gamma(n+\frac{1}{2}-\rho_m)} C_{n-k}^{(m)}(p) \\ &= \left\{ \sum_{k=1}^{n-m+1} \frac{n!(n-1)!(k+m-1)}{(k-1)!(k+2m)!} \frac{\Gamma\left(k+m-\frac{3}{2}+\rho_m\right)\Gamma\left(k+m-\frac{3}{2}-\rho_m\right)}{\Gamma\left(n+\frac{1}{2}+\rho_m\right)\Gamma(n+\frac{1}{2}-\rho_m)} \right\} \\ &\quad \times 4(2m+1)(-1)^{m+1}T_{m-1}^1(1-2p), \end{aligned} \tag{2.32}$$

where $\rho_m = \sqrt{m(m+1) + \rho + 1/4}$.

LEMMA 2.3.2. *For* $n = 1, 2, ...; m = 1, 2, ..., n,$

$$\sum_{k=1}^{n-m+1} \frac{n!(n-1)!(k+m-1)}{(k-1)!(k+2m)!} \frac{\Gamma\left(k+m-\frac{3}{2}+\rho_m\right)\Gamma\left(k+m-\frac{3}{2}-\rho_m\right)}{\Gamma\left(n+\frac{1}{2}+\rho_m\right)\Gamma\left(n+\frac{1}{2}-\rho_m\right)}$$
$$= \frac{n!(n-1)!}{(n+m-1)!(n-m)!} \frac{-1}{2(2m+1)} \left\{ \frac{1}{2m+\rho} + \frac{1}{2(m+1)-\rho} \frac{(n-m)(n-m-1)}{(n+m)(n+m+1)} \right\}. \tag{2.33}$$

PROOF. It is straightforward to check the identity for $m = n$. For $m = 1, 2, ..., n-1$, the finite series can be expressed as

$$\begin{aligned} &\sum_{k=1}^{n-m+1} \frac{n!(n-1)!(k+m-1)}{(k-1)!(k+2m)!} \frac{\Gamma\left(k+m-\frac{3}{2}+\rho_m\right)\Gamma\left(k+m-\frac{3}{2}-\rho_m\right)}{\Gamma\left(n+\frac{1}{2}+\rho_m\right)\Gamma\left(n+\frac{1}{2}-\rho_m\right)} \\ = \; & \frac{n!(n-1)!}{\Gamma\left(n+\frac{1}{2}+\rho_m\right)\Gamma\left(n+\frac{1}{2}-\rho_m\right)} \\ & \times \left[\frac{m\Gamma\left(m-\frac{1}{2}+\rho_m\right)\Gamma\left(m-\frac{1}{2}-\rho_m\right)}{(2m+1)!} y_{n-m}\left(m-\frac{1}{2}+\rho_m, m-\frac{1}{2}-\rho_m, 2m+2, 1\right) \right. \\ & \left. + \frac{\Gamma\left(m+\frac{1}{2}+\rho_m\right)\Gamma\left(m+\frac{1}{2}-\rho_m\right)}{(2m+2)!} y_{n-m-1}\left(m+\frac{1}{2}+\rho_m, m+\frac{1}{2}-\rho_m, 2m+3, 1\right) \right], \end{aligned} \tag{2.34}$$

where $y_n(a, b, c, z)$ is the truncated hypergeometric series (Erdélyi, 1953). The truncated hypergeometric series can be expressed as

$$y_i(a, b, c, 1) = \frac{\Gamma(a+i+1)\Gamma(b+i+1)}{i!\Gamma(a+b+i+1)} {}_3F_2\left(\begin{array}{c} a, b, c+i; 1 \\ c, a+b+i+1 \end{array} \right), \tag{2.35}$$

where

$${}_3F_2\left(\begin{array}{c} a, b, c; z \\ d, e \end{array} \right) \tag{2.36}$$

is the generalized hypergeometric series (Erdélyi, 1953). Thus, we have an identity for the truncated hypergeometric series:

$$\begin{aligned} y_i(a, b, a+b+j, 1) &= \frac{\Gamma(a+i+1)\Gamma(b+i+1)}{i!\Gamma(a+b+i+1)} {}_3F_2\left(\begin{array}{c} a, b, a+b+i+j; 1 \\ a+b+j, a+b+i+1 \end{array} \right) \\ &= \frac{\Gamma(a+i+1)\Gamma(b+i+1)}{i!\Gamma(a+b+i+1)} {}_3F_2\left(\begin{array}{c} a, b, a+b+i+j; 1 \\ a+b+i+1, a+b+j \end{array} \right) \\ &= \frac{\Gamma(a+i+1)\Gamma(b+i+1)}{i!\Gamma(a+b+i+1)} \frac{(j-1)!\Gamma(a+b+j)}{\Gamma(a+j)\Gamma(b+j)} \\ &\quad \times y_{j-1}(a, b, a+b+i+1, 1). \end{aligned} \tag{2.37}$$

By using the identity, we obtain

$$\begin{aligned}
&\sum_{k=1}^{n-m+1} \frac{n!(n-1)!(k+m-1)}{(k-1)!(k+2m)!} \frac{\Gamma\left(k+m-\frac{3}{2}+\rho_m\right)\Gamma\left(k+m-\frac{3}{2}-\rho_m\right)}{\Gamma\left(n+\frac{1}{2}+\rho_m\right)\Gamma\left(n+\frac{1}{2}-\rho_m\right)} \\
&= \frac{n!(n-1)!\Gamma\left(m-\frac{1}{2}+\rho_m\right)\Gamma\left(m-\frac{1}{2}-\rho_m\right)}{(n+m-1)!(n-m)!\Gamma\left(m+\frac{5}{2}+\rho_m\right)\Gamma\left(m+\frac{5}{2}-\rho_m\right)} \\
&\times \left\{ 2my_2\left(m-\frac{1}{2}+\rho_m, m-\frac{1}{2}-\rho_m, n+m, 1\right)\right. \\
&\left. + \frac{(n-m)\left(m-\frac{1}{2}+\rho_m\right)\left(m-\frac{1}{2}-\rho_m\right)}{n+m} y_1\left(m+\frac{1}{2}+\rho_m, m+\frac{1}{2}-\rho_m, n+m+1, 1\right)\right\} \\
&= \frac{n!(n-1)!}{(n+m-1)!(n-m)!} \frac{-1}{2(2m+1)} \left\{ \frac{1}{2m+\rho} + \frac{1}{2(m+1)-\rho} \frac{(n-m)(n-m-1)}{(n+m)(n+m+1)} \right\}.
\end{aligned} \tag{2.38}$$

□

By using Lemma 2.3.2, we have

$$\begin{aligned}
F_n^{(m)}(p) &= \frac{n!(n-1)!}{(n+m-1)!(n-m)!} \left\{ \frac{1}{2m+\rho} + \frac{1}{2(m+1)-\rho} \frac{(n-m)(n-m-1)}{(n+m)(n+m+1)} \right\} \\
&\quad \times 2(-1)^m T_{m-1}^1(1-2p).
\end{aligned} \tag{2.39}$$

By using (2.39) and the orthogonal property of the Gegenbauer polynomial, we have for $m = 2, 3, ...$

$$\begin{aligned}
&E_n^{(m)}(p,q,D) = (-1)^m \frac{n!(n-1)!}{(n+m)!(n-m-1)!} \\
&\times \left[\frac{2(2m+1)}{m(m+1)} p(1-p)q T_{m-1}^1(1-2p) + 2\left\{ \frac{T_m^1(1-2p)}{2(m+1)+\rho} + \frac{T_{m-2}^1(1-2p)}{2m-\rho} \right\} D \right],
\end{aligned} \tag{2.40}$$

and

$$E_n^{(1)}(p,q,D) = -3\frac{n-1}{n+1}\left\{ p(1-p)q + \frac{2(1-2p)}{4+\rho} D \right\}. \tag{2.41}$$

It is worthwhile to note that

$$\mu_{n,1,0}(t) \to \mu_{n,0,0}(t) \times q, \qquad \rho \to \infty. \tag{2.42}$$

The property agrees with the limit theorem given by Ethier (1979), where the three-dimensional diffusion process discussed here converges to the process which is the direct product of the one-dimensional processes for each locus.

By taking the limit $n \to \infty$, we have an expression for $\mu_{\infty,1,0}(t) = \lim_{n\to\infty} \mathbb{E}[x^n y]$ with

$$F_\infty^{(m)}(p) = \frac{4(2m+1)(-1)^m}{(2m+\rho)\{2(m+1)-\rho\}} T_{m-1}^1(1-2p) \tag{2.43}$$

and

$$E_{\infty}^{(m)}(p,q,D) = \frac{(2m+1)!}{m!(m+1)!} E_{m+1}^{(m)}(p,q,D), \tag{2.44}$$

and we arrive at an expression for (2.13):

$$\begin{aligned}
\mathbb{E}[x_1, x \in (0,1)] &= \frac{\rho D}{2+\rho} e^{-(1+\frac{\rho}{2})t} + 3\left\{ p(1-p)q + \frac{2(1-2p)}{4+\rho} D \right\} e^{-t} \\
&\quad - \sum_{m=2}^{\infty} 2(-1)^m \left[\frac{2m+1}{m(m+1)} pq(1-p) T_{m-1}^1(1-2p) \right. \\
&\quad \left. + \left\{ \frac{T_m^1(1-2p)}{2(m+1)+\rho} + \frac{T_{m-2}^1(1-2p)}{2m-\rho} \right\} D \right] e^{-\frac{m(m+1)}{2}t} \\
&\quad - \sum_{m=1}^{\infty} \frac{4(2m+1)(-1)^m}{(2m+\rho)\left[2(m+1)-\rho\right]} D T_{m-1}^1(1-2p) e^{-\frac{\rho+m(m+1)}{2}t}.
\end{aligned} \tag{2.45}$$

As $N \to \infty$, we observe $\mathbb{E}[x_1, x \in (0,1)] \to pq + De^{-ct}$, which shows the deterministic behavior of the gamete frequency x_1 without random genetic drift, as expected. We have the asymptotic form

$$\mathbb{E}[x_1, x \in (0,1)] \to 3\left\{ p(1-p)q + \frac{2(1-2p)}{4+\rho} D \right\} e^{-t}, \qquad t \to \infty. \tag{2.46}$$

The conditional expectation of the gamete frequency x_1 given that the locus A remains segregating is

$$\mathbb{E}[x_1 | x \in (0,1)] = \frac{\mathbb{E}[x_1, x \in (0,1)]}{\mathbb{P}[x \in (0,1)]}, \tag{2.47}$$

where the denominator is given by (2.1). The asymptotic form is

$$\mathbb{E}[x_1 | x \in (0,1)] \to \frac{q}{2} + \frac{(1-2p)D}{p(1-p)(4+\rho)}, \qquad t \to \infty. \tag{2.48}$$

In contrast to the deterministic model without random genetic drift, the value is higher than pq, to which the deterministic model tends. The conditional covariance between the frequencies of the alleles A_1 and B_1 is

$$Cov[x, y | x \in (0,1)] \to \frac{(1-2p)(1-q)}{(1-p)(4+\rho)}, \qquad t \to \infty. \tag{2.49}$$

In contrast to the deterministic model without random genetic drift, the finite value remains asymptotically. Moreover, the asymptotic value vanishes at $\rho \to \infty$, as is expected by the limit theorem by Ethier (1979).

The process of the change in the conditional expectation of the gamete frequency x_1 when the linkage disequilibrium is introduced into a population as $p = 1/2N = 0.05$ and $q = 0.2$ is illustrated in Figure 2.2. It can be seen that after $4N$ generations ($t = 2.0$) the conditional expectation of the gamete frequency x_1 almost reaches the asymptotic value

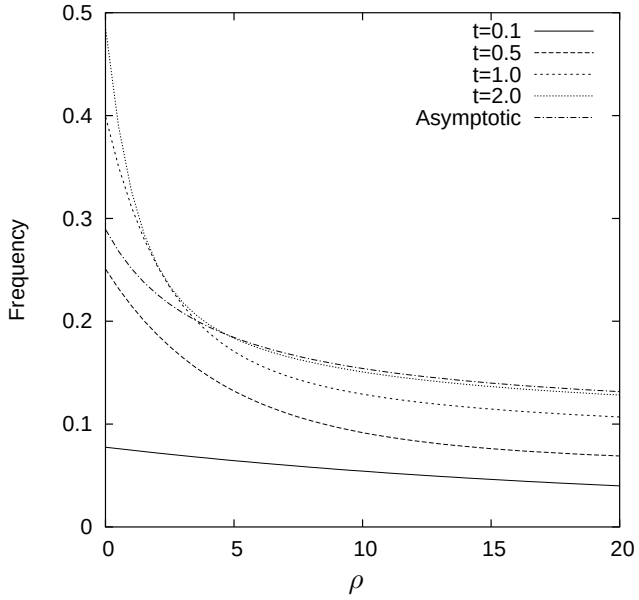

FIGURE 2.2. The conditional expectation of the gamete frequency x_1 given that the locus A remains segregating. $p = 0.05$ and $q = 0.2$.

for large ρ, although $4N$ generations is still not enough to reach the asymptotic value for small ρ. It can also be seen that the conditional expectation of the gamete frequency x_1 does not show monotonic behavior for small ρ. It increases rapidly and then decreases to the asymptotic value. For comparison, the counter part in the deterministic model is also illustrated in Figure 2.3.

2.4. Linkage disequilibrium in exponentially growing populations

The process generated by the generator (2.8) can be represented by a system of stochastic differential equations (Maruyama and Takahata, 1981; Maruyama, 1982). Let $\boldsymbol{B} = (B_1, B_2, B_3)'$ be a vector of independent Brownian motions. Here and subsequently $\boldsymbol{a}'$ denotes the transpose of matrix $\boldsymbol{a}$. The system of stochastic differential equations is

$$d\boldsymbol{x} = \boldsymbol{\sigma} d\boldsymbol{B} + \boldsymbol{v} dt, \tag{2.50}$$

where

$$\boldsymbol{x} = (x, y, z)', \qquad \boldsymbol{v} = \left(0, 0, -z\left(1 + \frac{\rho}{2}\right)\right)'. \tag{2.51}$$

σ is the square root of the covariance matrix of $\boldsymbol{x}$, whose explicit expression is given in Appendix. By using the Itô formula, we obtain a system of differential equations

$$\frac{d\boldsymbol{\mu}}{dt} = \boldsymbol{A}\boldsymbol{\mu}, \tag{2.52}$$

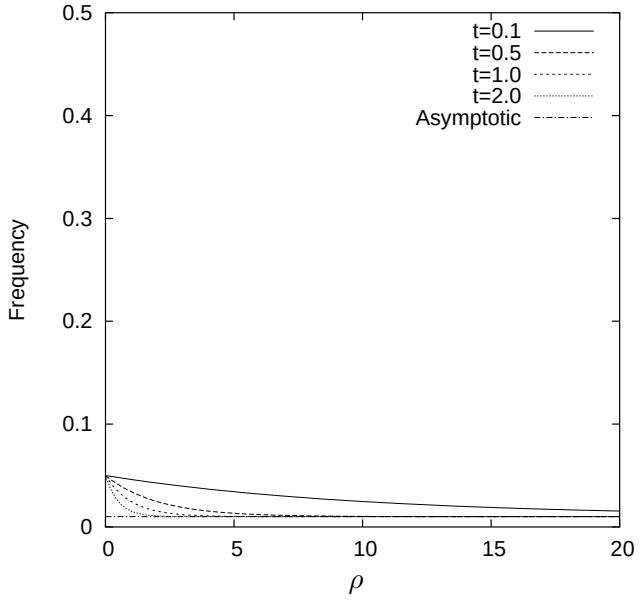

FIGURE 2.3. The gamete frequency x_1 in the deterministic model without random genetic drift. $p = 0.05$ and $q = 0.2$.

where

$$\boldsymbol{\mu}(t) = \left(\mathbb{E}[xy(1-x)(1-y)], \mathbb{E}[z(1-2x)(1-2y)], \mathbb{E}[z^2]\right)', \tag{2.53}$$

and

$$\boldsymbol{A} = \begin{pmatrix} -2 & 1 & 0 \\ 0 & -\left(5+\frac{\rho}{2}\right) & 4 \\ 1 & 1 & -(3+\rho) \end{pmatrix}. \tag{2.54}$$

The initial condition is

$$\boldsymbol{\mu}(0) = \left(pq(1-p)(1-q), D(1-2p)(1-2q), D^2\right)'. \tag{2.55}$$

The derivation is given in Appendix. The solution of the differential equation (2.52) is

$$\boldsymbol{\mu}(t) = e^{t\boldsymbol{A}}\boldsymbol{\mu}(0), \tag{2.56}$$

which reproduces the solution which was obtained by Ohta and Kimura (1969a). The elements of the matrix $e^{t\boldsymbol{A}}$ are given in Mano (2007). They involves three eigenvalues of the matrix $\boldsymbol{A}$. Denote them as $\lambda_i,\ i = 1, 2, 3$ with $0 > \lambda_1 > \lambda_2 > \lambda_3$. These eigenvalues satisfy a cubic equation

$$\lambda^3 + \left(10 + \frac{3}{2}\rho\right)\lambda^2 + \left(\frac{\rho^2}{2} + \frac{19}{2}\rho + 27\right)\lambda + \rho^2 + 13\rho + 18 = 0, \tag{2.57}$$

and

$$\begin{aligned}\lambda_1 &= -\frac{\rho}{2} - \frac{10}{3} + \frac{1}{3}\sqrt{76+6\rho+3\rho^2}\cos\frac{\varphi}{3},\\ \lambda_2 &= -\frac{\rho}{2} - \frac{10}{3} + \frac{1}{3}\sqrt{76+6\rho+3\rho^2}\cos\frac{\varphi+4\pi}{3},\\ \lambda_3 &= -\frac{\rho}{2} - \frac{10}{3} + \frac{1}{3}\sqrt{76+6\rho+3\rho^2}\cos\frac{\varphi+2\pi}{3},\end{aligned}$$

where φ $(0 < \varphi < \pi)$ satisfies

$$\cos\varphi = -\frac{224+126\rho-45\rho^2}{(76+6\rho+3\rho^2)^{\frac{3}{2}}}. \tag{2.58}$$

They are

(2.59)

$$\lambda_1 = -2 + \frac{8}{\rho^2} + O(\rho^{-3}), \qquad \lambda_2 = -\frac{\rho}{2} - 5 + \frac{8}{\rho} + O(\rho^{-2}), \qquad \lambda_3 = -\rho - 3 - \frac{8}{\rho} + O(\rho^{-2}).$$

Note that λ_i here are twice of those in Ohta and Kimura (1969a). Figure 1 in Ohta and Kimura (1969a) plots the halves of these eigenvalues as functions of ρ.

Next, consider a random mating diploid population with an initial effective size of N and where the effective size changes from generation to generation in a deterministic way. The Wright-Fisher model is time-inhomogeneous, since the effective size depends on time. Assume that, as for the diffusion model, the effective size is sufficiently large in the time period such that the gamete frequencies can be regarded as continuous variables. Also, assume that the effective size grows continuously such that it can be represented as a continuous function of time s (measured in units of one generation). To this end, define the relative function $\lambda(s)$ by

$$\lambda_N(s) = \frac{N}{N(\lceil s \rceil)} = \frac{N}{N(j)}, \qquad j-1 < s \le j, \qquad j = 1, 2, ..., \tag{2.60}$$

and $N(0) = N, \lambda_N(0) = 1$. We are interested in the behavior of the process $\lim_{N\to\infty}\lambda_N(s) = \lambda(s)$. To avoid confusions, we will show time dependence of $N(s)$.

$$\tau = \int_0^s \frac{du}{2N(u)} = \frac{\Lambda(s)-\Lambda(0)}{2N}, \tag{2.61}$$

where $\Lambda(s)$ is a primitive function of $\lambda(s)$. Note that τ is a time measured in units of harmonic mean of twice of the population sizes between 0 and s. This model is the time-inhomogeneous diffusion process $\{x(\tau), y(\tau), z(\tau) : \tau_\infty \ge \tau \ge 0\}$, where

$$\tau_\infty = \int_0^\infty \frac{du}{2N(u)}, \tag{2.62}$$

in the same three dimensional domain as the diffusion model for constant size populations. The time-inhomogeneous diffusion process is represented by a system of stochastic

differential equations by replacing N with $N(s)$ in the system of stochastic differential equations (2.50), and we have

$$d\boldsymbol{x} = \boldsymbol{\sigma} d\boldsymbol{B} + \boldsymbol{v}(\tau)d\tau, \tag{2.63}$$

where

$$\boldsymbol{v}(\tau) = (0, 0, -z(1 + \rho(\tau)/2)' \tag{2.64}$$

and $\rho(\tau) = 4N(s)r$. For a population exponentially growing at a rate e^b $(b > 0)$ times per generation, we have $\lambda(s) = e^{-bs}$, and

$$\tau = \frac{1 - e^{-\beta t}}{\beta}, \qquad \tau_\infty = \frac{1}{\beta}, \tag{2.65}$$

where $\beta = 2Nb$. A diffusion time scaling is to let $2N \to \infty$, while β and ρ are hold constant. By applying the Itô formula, we obtain a system of differential equations:

$$\frac{d\boldsymbol{\mu}}{d\tau} = \boldsymbol{A}\boldsymbol{\mu} - \frac{\rho}{2}\frac{\beta\tau}{1 - \beta\tau}\boldsymbol{C}\boldsymbol{\mu}, \qquad \boldsymbol{C} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}. \tag{2.66}$$

This differential equation can be solved numerically, although it is difficult to solve explicitly. The Maclaurin expansion of the second term of the right hand side of (2.66) around $\beta\tau = 0$, which converges $\beta\tau < 1$, gives

$$\frac{d\boldsymbol{\mu}}{d\tau} = \boldsymbol{A}\boldsymbol{\mu} - \frac{\rho}{2}\sum_{n=1}^{\infty}(\beta\tau)^n\boldsymbol{C}\boldsymbol{\mu}. \tag{2.67}$$

When the growth rate is not large such that the solution can be well expressed by a perturbative series in β with few terms, we have

$$\boldsymbol{\mu}(\tau) = \sum_{n=0}^{\infty}\beta^n\boldsymbol{\mu}^{(n)}(\tau), \tag{2.68}$$

where $\boldsymbol{\mu}^{(0)}(\tau)$ is given by (2.56), and

$$\boldsymbol{\mu}^{(n)}(\tau) = -\frac{\rho_0}{2}\sum_{i=0}^{n-1}\int_0^\tau \zeta^{n-i}e^{(\tau-\zeta)\boldsymbol{A}}\boldsymbol{C}\boldsymbol{\mu}^{(i)}(\zeta)d\zeta = \sum_{i=0}^{2n}\sum_{j=1}^{3}\boldsymbol{\mu}_{j,i}^{(n)}\tau^i e^{\lambda_j\tau}, \tag{2.69}$$

for $n \geq 1$. The perturbative series is parametric in τ (Hinch, 1991). A system of recurrence relations for $\boldsymbol{\mu}_{j,i}^{(n)}$ are given in Mano (2007). The exact convergence radius of the perturbative series is not known. Convergence of a perturbative series is not necessary, since we will always use a truncated series of finite terms; What we have to know is how a truncated series approximates the exact solution for each fixed parameter (Hinch, 1991). The error of the truncated perturbative series was examined by using the exact solutions

obtained by numerical integration of (2.66) and the perturbative series truncated at the 10th order term. The error $|\boldsymbol{\mu}_{10}(\tau)/\boldsymbol{\mu}(\tau) - 1|$, where $\boldsymbol{\mu}_{10}(\tau)$ is the truncated series, was less than 0.01 when $\beta < 0.85/\tau$ and $\beta < 20$. When the growth rate is very large the perturbative analysis above is useless. However, rapid size growth never continue for a long period in natural populations. The rapid size growth can be modeled by a stepwise growth (Slatkin, 1994a). It is straightforward to obtain the solution by connecting two solutions to the periods before and after the size growth.

Let us consider the asymptotic behavior of the squared standard linkage deviation for exponentially growing populations after a large number of generations. For a constant size population with an effective size of N, we have (2.10). Assume the growth rate is not large. Let us consider the asymptotic behavior of the squared standard linkage deviation for large t ($< \infty$). In fact, $t = \infty$ cannot be considered here, because the infinite series in (2.67) does not converge. Note that

$$\tau = -\frac{1}{\beta}\sum_{i=1}^{\infty}\frac{(-\beta t)^i}{i!} = t - 2\beta t^2 + O(\beta^2). \tag{2.70}$$

Suppose τ is sufficiently large such that the asymptotic rates of decrease of the three moments (2.53) are given by $\lambda_1\,(< 0)$, which is the largest eigenvalue of the matrix $\boldsymbol{A}$. Here, $\lambda_1 - \lambda_2$ and $\lambda_1 - \lambda_3$ are $O(\rho)$. (2.68) is approximated as

$$\boldsymbol{\mu}(\tau) \approx \sum_{n=0}^{\infty}\sum_{i=0}^{2n}\beta^n \boldsymbol{\mu}_{1,i}^{(n)}\tau^i e^{\lambda_1\tau}, \qquad t \to \infty, \tag{2.71}$$

and up to the first order perturbation, we have

$$\boldsymbol{\mu}(\tau) \approx \left\{\boldsymbol{\mu}_{1,0}^{(0)} + \left(\boldsymbol{\mu}_{1,0}^{(1)} + \boldsymbol{\mu}_{1,1}^{(1)}\tau + \boldsymbol{\mu}_{1,2}^{(1)}\tau^2\right)\beta\right\} e^{\lambda_1\tau}, \qquad t \to \infty, \tag{2.72}$$

where explicit expressions for $\boldsymbol{\mu}_{1,0}^{(0)}$ and $\boldsymbol{\mu}_{1,i}^{(1)}$ are given in Mano (2007). The standard linkage deviation tends to be

$$\begin{aligned}\sigma_d^2(\tau) \;\approx\; & \frac{\mu_{3,1,0}^{(0)}}{\mu_{1,1,0}^{(0)}} + \left\{\left(\frac{\mu_{3,1,0}^{(1)}}{\mu_{1,1,0}^{(0)}} - \frac{\mu_{3,1,0}^{(0)}\mu_{1,1,0}^{(1)}}{\mu_{1,1,0}^{(0)\,2}}\right) + \left(\frac{\mu_{3,1,1}^{(1)}}{\mu_{1,1,0}^{(0)}} - \frac{\mu_{3,1,0}^{(0)}\mu_{1,1,1}^{(1)}}{\mu_{1,1,0}^{(0)\,2}}\right)\tau \right. \\ & \left. + \left(\frac{\mu_{3,1,2}^{(1)}}{\mu_{1,1,0}^{(0)}} - \frac{\mu_{3,1,0}^{(0)}\mu_{1,1,2}^{(1)}}{\mu_{1,1,0}^{(0)\,2}}\right)\tau^2\right\}\beta, \qquad t \to \infty.\end{aligned} \tag{2.73}$$

For large ρ, we obtain the asymptotic formula

$$\sigma_d^2(\tau) \;\approx\; \frac{1}{\rho} - \frac{\tau}{\rho}\beta + O\left(\rho^{-2}\right) = \frac{1}{\rho(s)} + O\left(\rho^{-2}\right), \qquad t \to \infty. \tag{2.74}$$

The asymptotic value is determined by the current (at s-th generation) effective size of the population and the recombination rate. It is dependent neither the initial condition for the moments, the initial effective size, nor the growth rate.

2.5. Stationary state with mutations

For constant size populations, the asymptotic formula for the squared standard linkage deviation (2.10) is universal; it holds for various types of models with mutations and natural selections (Kimura and Ohta, 1971). It holds irrespective of whether a state is in steady decay without mutations or in equilibrium with mutation. Let us investigate whether the universality holds in exponentially growing populations. The recurrent mutation model is assumed, whereby mutations occur between two types of alleles (Ohta and Kimura, 1969b). For simplicity, let the mutation rate per generation per locus between the alleles A_1 and A_2 and that between the alleles B_1 and B_2 be u,which is usually smaller than c. As noted above, for a constant size population with an effective size of N, it was shown that the asymptotic formula (2.10) holds in the mutation-drift equilibrium (Ohta and Kimura, 1969b).

For the model, the system of stochastic differential equation is

$$d\boldsymbol{x} = \boldsymbol{\sigma} d\boldsymbol{B} + \boldsymbol{v}_m(\tau)d\tau, \tag{2.75}$$

where

$$\boldsymbol{v}_m(\tau) = ((1-2x)\theta(\tau)/2, (1-2y)\theta(\tau)/2, -z(1+2\theta(\tau)+\rho(\tau)/2))' \tag{2.76}$$

and $\theta(\tau) = 4N(\tau)u$. A diffusion time scaling is to measure time in units of $2N$ generations and let $2N \to \infty$, while $\theta = 4Nu, \beta$ and ρ are hold constant. By applying the Itô formula, we obtain a system of differential equations

$$\frac{d\boldsymbol{\mu}}{d\tau} = (\boldsymbol{A} - 4\theta\boldsymbol{E})\boldsymbol{\mu} - \frac{\beta\tau}{1-\beta\tau}\left(\frac{\rho}{2}\boldsymbol{C} + 4\theta\boldsymbol{E}\right)\boldsymbol{\mu} + \left(\frac{\theta(\tau)}{2}\mu_4(\tau), 0, 0\right)', \tag{2.77}$$

where $\mu_4(\tau) = \mathbb{E}[x(1-x) + y(1-y)]$, $\boldsymbol{E}$ is the identity matrix. The moment μ_4 satisfies the differential equation

$$\frac{d\mu_4}{d\tau} = \theta(\tau) - \{1 + 2\theta(\tau)\}\,\mu_4 \tag{2.78}$$

and it is straightforward to solve it and

$$\begin{aligned} \mu_4(\tau) &= [p(1-p) + q(1-q)](1-\beta\tau)^{2\theta/\beta}e^{-\tau} \\ &+ \frac{\theta(1-\beta\tau)^{2\theta/\beta}}{\beta^{1+2\theta/\beta}}\left\{\Psi\left(1+\frac{2\theta}{\beta}, 1+\frac{2\theta}{\beta}; \frac{1}{\beta}-\tau\right) - e^{-\tau}\Psi\left(1+\frac{2\theta}{\beta}, 1+\frac{2\theta}{\beta}; \frac{1}{\beta}\right)\right\}, \end{aligned} \tag{2.79}$$

where $\Psi(a,b;z)$ represents a confluent hypergeometric function (Erdélyi, 1953). The system of differential equations (2.77) can be solved numerically, although it is difficult to solve explicitly. Assume the growth rate is not large such that the solution can be well expressed by a perturbative series in β with few terms. Up to the first order perturbation, we have

$$\begin{aligned}\mu_4(\tau) &\approx \frac{\theta}{1+2\theta} + \left\{p(1-p)+q(1-q)-\frac{\theta}{1+2\theta}\right\} e^{-(1+2\theta)\tau} \\ &\quad + \left[\left\{\frac{\theta\tau}{(1+2\theta)^2} + \frac{\theta(e^{-(1+2\theta)\tau}-1)}{(1+2\theta)^3}\right\} - 2\theta_2\tau^2 e^{-(1+2\theta)\tau}\right]\beta, \end{aligned} \tag{2.80}$$

where

$$\theta_1 = \frac{\theta^2}{2(1+2\theta)}, \qquad \theta_2 = \frac{\theta}{2}\{p(1-p)+q(1-q)\} - \theta_1. \tag{2.81}$$

By substituting the expression into the system of differential equations (2.77), we obtain the zeroth order solution

$$\begin{aligned}\boldsymbol{\mu}^{(0)}(\tau) &= e^{\tau(\boldsymbol{A}-4\theta\boldsymbol{E})}\boldsymbol{\mu}(0) + \int_0^\tau e^{(\tau-\zeta)(\boldsymbol{A}-4\theta\boldsymbol{E})}\left(\frac{\theta}{2}\mu_4^{(0)}(\zeta),0,0\right)' d\zeta \\ &= \sum_{j=1}^3 \boldsymbol{\mu}_{j,0}^{(0)} e^{(\lambda_j-4\theta)\tau} - \sum_{j=1}^3 \boldsymbol{\eta}_j^{(0)}\left[\frac{\theta_1}{\lambda_j-4\theta} + \frac{\theta_2}{\lambda_j+1-2\theta}\left\{e^{-(1+2\theta)\tau} - e^{(\lambda_j-4\theta)\tau}\right\}\right], \end{aligned} \tag{2.82}$$

where explicit expressions for $\boldsymbol{\mu}_{j,0}^{(0)}$ are given in Mano (2007), and $(e^{t\boldsymbol{A}})_{i1} = \sum_{j=1}^3 (\boldsymbol{\eta}_j^{(0)})_i e^{\lambda_j t}$. The asymptotic values

$$\boldsymbol{\mu}^{(0)}(\infty) = -\sum_{j=1}^3 \frac{\boldsymbol{\eta}_j^{(0)}\theta_1}{\lambda_j - 4\theta} \tag{2.83}$$

give the values where a constant size population with an effective size of N is in the mutation-drift equilibrium. This is the same as Equations 17 of Ohta and Kimura (1969b), where they computed

$$\boldsymbol{\mu}^{(0)}(\infty) = -(\boldsymbol{A}-4\theta_0\boldsymbol{E})^{-1}(\theta_1,0,0)' \tag{2.84}$$

in our notations. It can be shown that these two expressions (2.83) and (2.84) are equivalent by noting that $\boldsymbol{\eta}_j^{(0)}$ are eigenvectors of λ_j. In the first order, we have

$$\begin{aligned}\boldsymbol{\mu}^{(1)}(\tau) &= -\int_0^\tau \zeta e^{(\tau-\zeta)(\boldsymbol{A}-4\theta\boldsymbol{E})}\left(\frac{\rho}{2}\boldsymbol{C}+4\theta\boldsymbol{E}\right)\boldsymbol{\mu}^{(0)}(\zeta)d\zeta \\ &\quad + \int_0^\tau e^{(\tau-\zeta)(\boldsymbol{A}-4\theta_0\boldsymbol{E})}\left(\frac{\theta_0}{2}(\mu_4^{(1)}(\zeta)+\zeta\mu_4^{(0)}(\zeta)),0,0\right)' d\zeta \\ &= \sum_{j=1}^3\left[\boldsymbol{\eta}_j^{(1)}\left\{\frac{\tau}{\lambda_j-4\theta}+\frac{1}{(\lambda_j-4\theta)^2}\right\} + \boldsymbol{\eta}_j^{(0)}\left\{\frac{\theta_3-\theta_4\tau}{\lambda_j-4\theta} - \frac{\theta_4}{(\lambda_j-4\theta)^2}\right\}\right] + \boldsymbol{\xi}(\tau), \end{aligned} \tag{2.85}$$

where

$$\theta_3 = \frac{\theta_1}{(1+2\theta)^2}, \qquad \theta_4 = \theta_1 + \frac{\theta_1}{1+2\theta}. \tag{2.86}$$

The explicit expressions for $\boldsymbol{\eta}^{(1)}$ are given in Mano (2007). The vector of functions $\boldsymbol{\xi}(\tau)$ represents terms that decay exponentially in τ.

Contrasting with the model without mutations, the three moments (2.53) do not vanish asymptotically after a large number of generations. The asymptotic values are given by terms which do not decay exponentially in τ. Up to the first order perturbation, we have

$$\boldsymbol{\mu}(\tau) \approx \boldsymbol{\mu}^{(0)}(\infty) + \sum_{j=1}^{3} \left[\boldsymbol{\eta}_j^{(1)} \left\{ \frac{\tau}{\lambda_j - 4\theta} + \frac{1}{(\lambda_j - 4\theta)^2} \right\} + \boldsymbol{\eta}_j^{(0)} \left\{ \frac{\theta_3 - \theta_4 \tau}{\lambda_j - 4\theta} - \frac{\theta_4}{(\lambda_j - 4\theta)^2} \right\} \right] \beta. \tag{2.87}$$

The standard linkage deviation tends to be

$$\begin{aligned} \sigma_d^2(\tau) \quad \approx \quad & \frac{\mu_3^{(0)}(\infty)}{\mu_1^{(0)}(\infty)} + \sum_{j=1}^{3} \left[\left\{ \frac{\eta_{3,j}^{(1)}}{\mu_1^{(0)}(\infty)} - \frac{\eta_{1,j}^{(1)} \mu_3^{(0)}(\infty)}{\mu_1^{(0)\,2}(\infty)} \right\} \left\{ \frac{\tau}{\lambda_j - 4\theta} + \frac{1}{(\lambda_j - 4\theta)^2} \right\} \right. \\ & \left. + \left\{ \frac{\eta_{3,j}^{(0)}}{\mu_1^{(0)}(\infty)} - \frac{\eta_{1,j}^{(0)} \mu_3^{(0)}(\infty)}{\mu_1^{(0)\,2}(\infty)} \right\} \left\{ \frac{\theta_3 - \theta_4 \tau}{\lambda_j - 4\theta} - \frac{\theta_4}{(\lambda_j - 4\theta)^2} \right\} \right] \beta, \qquad t \to \infty. \end{aligned} \tag{2.88}$$

For large ρ, we obtain an asymptotic formula which is exactly the same as the formula (2.74). Thus, the asymptotic formula holds regardless of the mutations.

2.6. Numerical examples and simulation results

Let us discuss the effects of population size growth on the squared standard linkage deviations by numerical examples. The differential equations (2.66,2.77) were solved numerically with the MATHEMATICA program (Wolfram Research, Inc. 2004). Two models without mutations after populations were founded were assumed, whereby linkage disequilibrium was introduced in the founder populations with an effective size of N. One model is the admixture model (Chakraborty and Weiss, 1988), whereby linkage disequilibrium is introduced by an admixture of the two equally sized populations where one population has the gamete A_1B_1 only and the other population has the gamete A_2B_2 only. This model specifies $p = q = 1/2, D = 1/4$. The other model is the single mutation model (Nei and Li, 1980), whereby linkage disequilibrium is introduced by a single mutation, as was discussed in Section 2.3. We assume that the locus A is not segregating and the wild type allele A_2, and the locus B, in which a pair of alleles B_1 and B_2 are evenly segregating. Then, the mutation introduces the mutant allele A_1 to the locus A of one of the allele B_1 bearing chromosomes. This model specifies $p = 1/(2N), q = 1/2, D = 1/(4N)$. For

the recurrent mutation model, the mutation-drift equilibrium was assumed for the founder populations with an effective size of N. The initial conditions for the three moments (2.53) were set to their equilibrium values (2.83) with $u = 10^{-5}$.

Accuracy of the asymptotic formula $\sigma_d^2 \approx 1/\rho(s)$ for the squared standard linkage deviation is shown in Figure 2.4. It was assumed that the populations with current effective sizes of $N(s) = 10,000$ were founded $s = 40,000$ generations ago with effective sizes of $N = 500, 1,000$, and 5,000. Figure 2.4 (a) shows the results for the admixture model, however, the results for the single mutation model were the same up to the 10^{-5} order. It can be seen that the asymptotic formula is highly accurate. Figure 2.4 (b) displays the results for the recurrent mutation model. It can be seen that the asymptotic formula is not accurate when ρ is small. The reason for the poor fitting is that the asymptotic formula is not applicable for small ρ. It is due to the mutation rate v being comparable to the recombination rate r, exemplifying the fact that mutations become the dominant factor for the decay of linkage disequilibrium between very closely linked loci.

The squared standard linkage deviation in populations with a current effective size of $N(s) = 10,000$ and founded $s = 4,000$ generations ago with effective sizes of $N = 500, 1,000$, and 5,000 are shown in Figure 2.5. Figure 2.5 (a), (b), and (c) show the results for the single mutation model, admixture model, and the recurrent mutation model, respectively. In contrast to Figure 2.4, it can be seen that the asymptotic formula $\sigma_d^2 \approx 1/\rho(s)$ does not give accurate values. Note that the perturbation is valid, since the perturbative series truncated at the 10th order term had small error. Compared with Figure 2.4, the poor fitting shows that the time period of the growth $s = 4,000$ is too short to apply the asymptotic formula.

The squared correlation coefficient of gamete frequencies (Hill and Robertson, 1968)

$$r^2 := \frac{z^2}{x(1-x)y(1-y)}, \tag{2.89}$$

has been used as a practical summary statistic for linkage disequilibrium. Unfortunately, analytical treatment of the statistics is not available. It has been discussed whether or not the squared standard linkage deviation approximates the expectation of the squared correlation coefficient. Maruyama (1982) used numerical integration of the stochastic differential equation to estimate expectation of the squared correlation coefficient. He assumed the infinite allele mutation model. He showed that the squared standard linkage deviation can differ substantially from the squared correlation coefficient. By simulations, Hudson (1985) showed that the squared standard linkage deviation gives a good approximation of

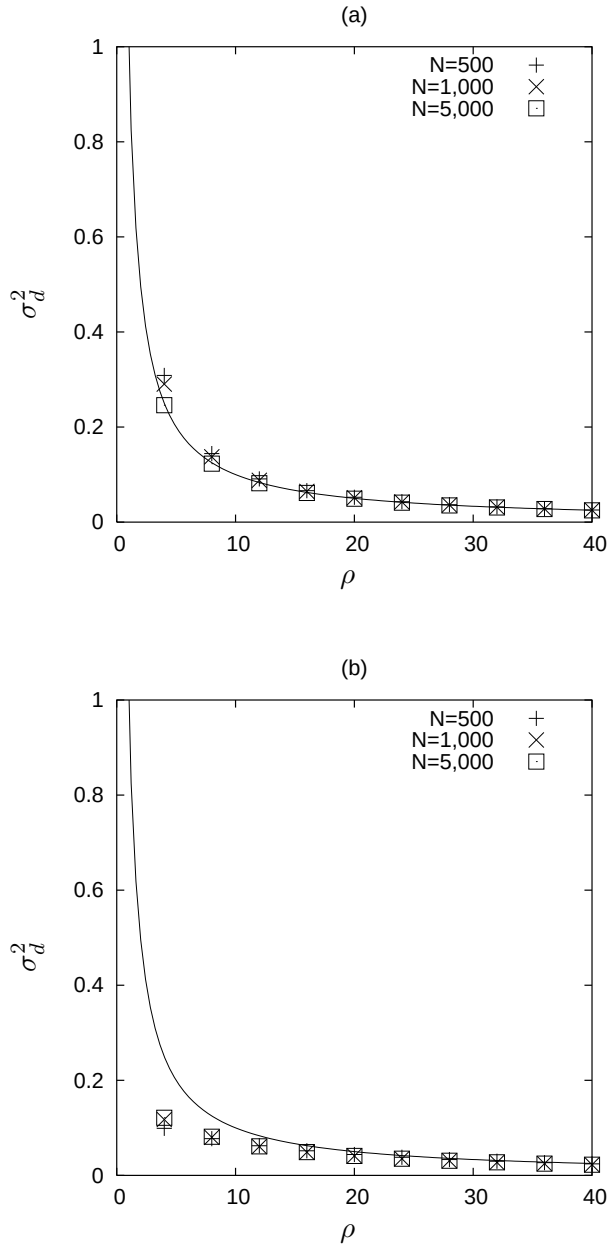

FIGURE 2.4. The squared standard linkage deviations in exponentially growing populations. $N(s) = 10,000$ and $s = 40,000$. (a) the admixture model; (b) the recurrent mutation model. The line shows values given by the asymptotic formula.

the expectation of the squared correlation coefficient, conditional on the minor allele frequencies being larger than 0.05. It is worthwhile examining whether the squared standard linkage deviation gives a good approximation of the expectation of the squared correlation coefficient in an exponentially growing population. Also, it is not trivial that the

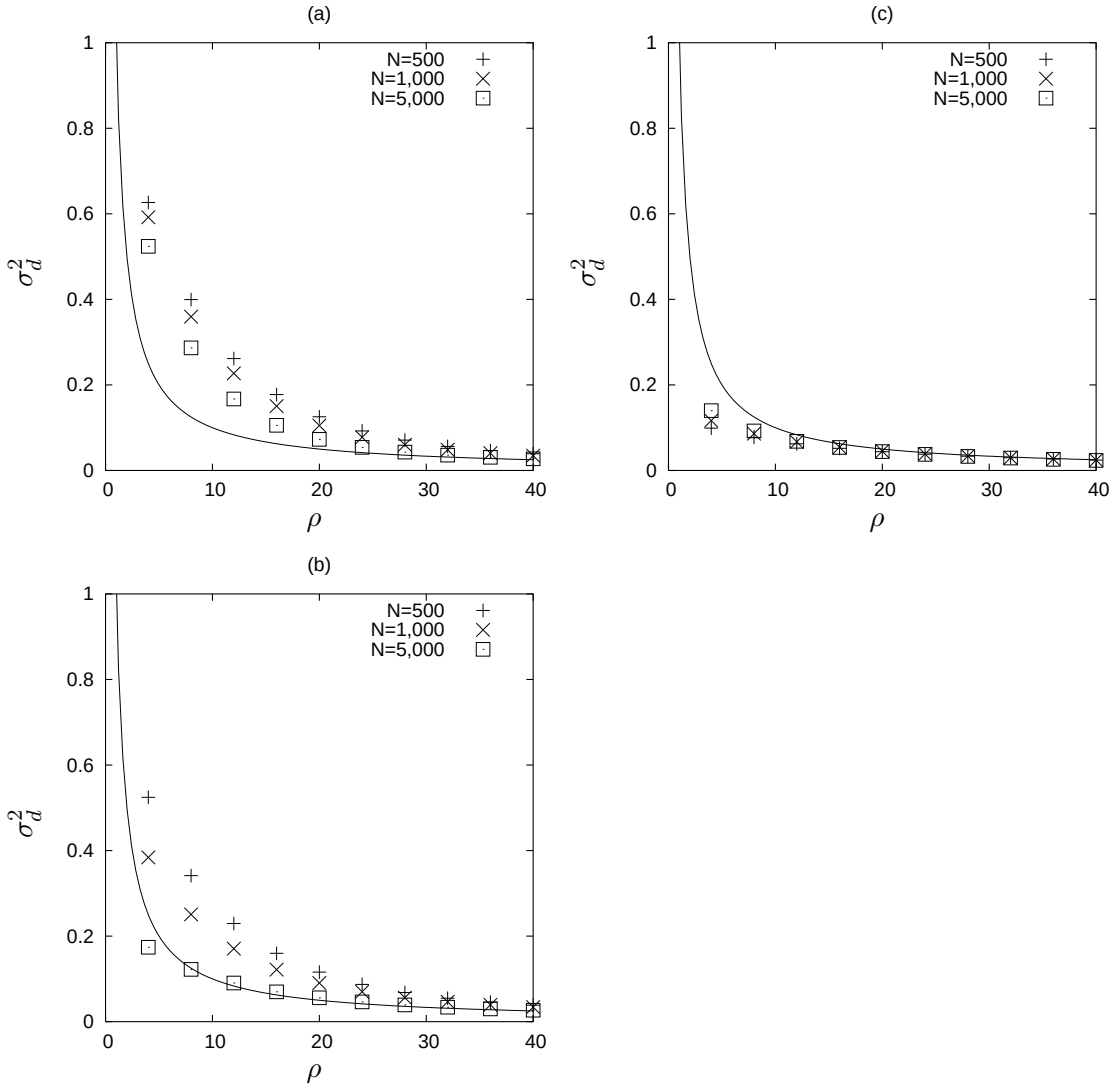

FIGURE 2.5. The squared standard linkage deviations in exponentially growing populations. $N(s) = 10,000$ and $s = 4,000$. (a) the admixture model; (b) the single mutation model; (c) the recurrent mutation model. The lines show values given by the asymptotic formula.

time-inhomogeneous diffusion process introduced here gives a good approximation of the time-inhomogeneous Wright-Fisher model. Thus, simulations of the time-inhomogeneous Wright-Fisher model were conducted for an exponentially growing population with the exponentially size growth. The single mutation model, the admixture model, and the recurrent mutation model described in Section 2.6 were assumed. The expectations of the squared correlation coefficient and the moments (2.53) were estimated by the arithmetic mean of the values in the simulated populations in which both of the loci were segregating.

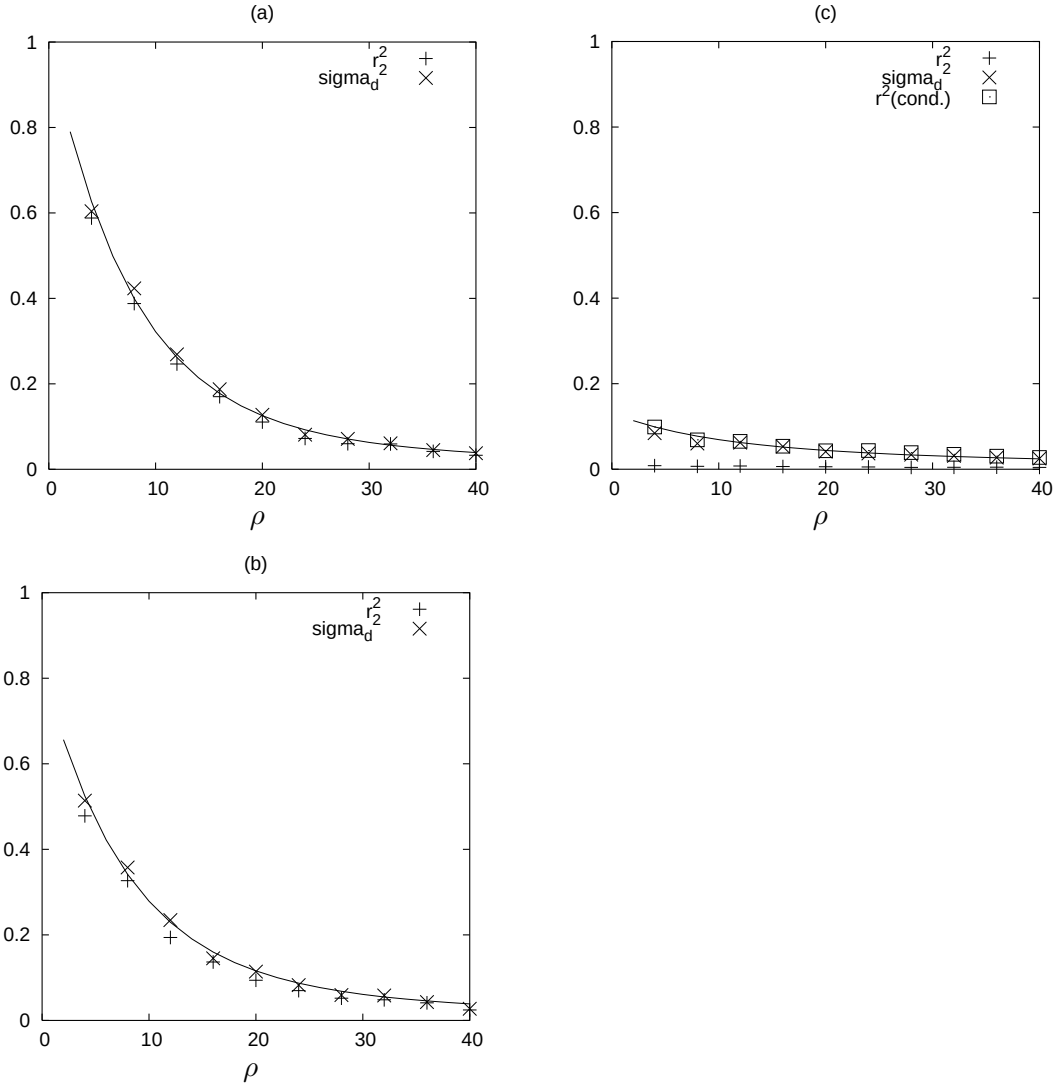

FIGURE 2.6. Simulation estimates for the squared standard linkage deviations ($\times$) and the expectation of the squared correlation coefficients ($+$) in exponentially growing populations. $N(s) = 10,000, s = 4,000$ and $N = 500$. The lines show the squared standard linkage deviations given by the time-inhomogeneous diffusion model. (a) the admixture model; (b) the single mutation model; (c) the recurrent mutation model. For the recurrent mutation model, simulation estimate for the squared correlation coefficient, conditional on the minor allele frequencies being larger than 0.05 ($\square$) is also shown.

The expectation of the squared correlation coefficient and the squared standard linkage deviation obtained by the simulations are shown in Figure 2.6. It was assumed that the

populations with a current effective size of $N(s) = 10,000$ were founded $s = 4,000$ generations ago with an effective size of $N = 500$. The results for the single mutation model and the admixture model are shown in Figure 2.6 (a), and (b), respectively. For the single mutation model and the admixture model, 100,000 and 1,000 populations were generated, respectively. The squared standard linkage deviation obtained by the simulations was close to that obtained by the solution to the system of differential equations (2.66), which means that the time-inhomogeneous diffusion model gives a good approximation of the time-inhomogeneous Wright-Fisher model. Also, it can be seen that the squared standard linkage deviation gives accurate approximation of the expectation of the squared correlation coefficient. The results for the recurrent mutation model are shown in Figure 2.6 (c). To generate founder populations in mutation-drift equilibrium with effective sizes of $N = 500$, $40,000$ generations were simulated. For the model, 10,000 populations were generated. It can be seen that the squared standard linkage deviation obtained by the simulations is close to that obtained by the solution to the system of differential equations (2.77). The squared standard linkage deviation was substantially smaller than the expectation of the squared correlation coefficient. However, the squared standard linkage deviation gave a fairly accurate approximation of the squared correlation coefficient, conditional on the minor allele frequencies being larger than 0.05, as was observed by Hudson (1985) for constant size populations.

2.7. Summary

The analytic expression of conditional expectation of transient gamete frequency given that one of the two loci remains segregating was obtained in terms of the diffusion process by calculating the moments of the distribution. This expression is independent of models which introduce linkage disequilibrium into a population. We considered the model that linkage disequilibrium is introduced by a single mutation and association between the mutant allele A_1 and the allele B_1, which filled the other locus of the chromosome on which the mutation occurred. Because the allele A_1 is prone to be lost by random genetic drift, the conditional expectation of the frequency of the gamete A_1B_1 given that the locus A remains segregating would describe our observation. The behavior is significantly different from the monotonic decrease in the deterministic model without random genetic drift. After $4N$ generations, the conditional expectation of the gamete frequency almost reaches the asymptotic value for large ρ, although $4N$ generations is still not enough for small ρ. The conditional covariance between the frequency of the alleles A_1 and

B_1 does not vanish asymptotically and it depends on the recombination rate between the loci. Note that the conditional expectation of the linkage disequilibrium coefficient monotonically decreases in a similar manner to that in the deterministic model (see 2.9). This observation demonstrates the obvious fact that the linkage disequilibrium measure is not enough to characterize the two-locus problem uniquely.

Then, evolution of linkage disequilibrium of the founders in exponentially growing populations was studied using a time-inhomogeneous diffusion model. By simulations, it was shown that the model is a good approximation of a time-inhomogeneous Wright-Fisher model, where the population size grows exponentially in a deterministic way. It is easy to extend this model to other types of demographic models. It was demonstrated that the squared linkage deviation can be obtained by solving a system of differential equations numerically. In addition, a perturbative solution was obtained when the growth rate is not large. By using the first order perturbation, an asymptotic formula for the squared standard linkage deviation after a large number of generations was obtained, although such long term growth, say, the number of generations same as the current effective size, may hardly occurred in natural populations. According to the formula, the squared standard linkage deviation tends to be $1/\rho(s)$. It is independent from either the initial effective size, the growth rate, or the mutation rate. It was shown that the squared standard linkage deviation gives a good approximation of the expectation of the squared correlation coefficient for models without mutation after populations are founded. For a model with mutation after populations are founded, the squared standard linkage deviation gave a good approximation of the expectation of the squared correlation coefficient, conditional on the minor allele frequencies being larger than 0.05. It is clear that the conjecture by Slatkin (1994a) does not hold. When a rapidly growing population is founded by small size in which there is already linkage disequilibrium between a particular pair of loci, the linkage disequilibrium decays faster than a constant size population whose effective size is the harmonic mean of the individual effective sizes over the time period of the growth.

Rogers and Harpending (1992) obtained an approximate distribution of the number of nucleotide site differences in pairwise comparisons of DNA sequences in growing populations. They showed that the size growth can be inferred by fitting the distribution to sample data. A differential equation for the expectation of the average number of nucleotide sites differences in pairwise comparisons of DNA sequences, or the nucleotide

diversity, in a size changing population is known (Li, 1977; Tajima, 1989b). In exponentially growing populations,

$$\mathbb{E}[\pi] = \pi(0)e^{-\tau} + \frac{\theta}{\beta}\left\{\Psi\left(1,1;\frac{1}{\beta}-\tau\right) - e^{-\tau}\Psi\left(1,1;\frac{1}{\beta}\right)\right\}. \tag{2.90}$$

Since $\Psi(1,1;x) \approx -\log x$ as $x \to 0$, after a large number of generations we have $\mathbb{E}[\pi] \approx \theta t = 2us$ irrespective of the population size. This expectation is consistent with star shape genealogies which are expected to be observed in growing populations, where all coalescence events occur at around the most common recent ancestor (Slatkin and Hudson, 1991). If we ignore mutations after the populations were founded, the second term in (2.90) can be dropped. Since τ is the harmonic mean of the individual effective sizes over the time period of the growth, the decay of the nucleotide diversity is similar to that in a constant size population whose effective size is the harmonic mean of the individual effective sizes over the time period of the growth. In contrast, the asymptotic formula $1/\rho(s)$ means that linkage disequilibrium in exponentially growing populations is asymptotically the same as that in a constant size population, the effective size of which is the current effective size. When we use the formula for constant size populations, the estimates of effective population size via linkage disequilibrium and those via the nucleotide diversity will give different estimates. Nevertheless, several authors have implicitly assumed that these two estimates can be equated (Pritchard and Przeworski, 2001; Przeworski and Wall, 2001). The expected squared standard linkage deviation in the current entire human population for the estimates $N = 3,200, N(s) = 550,000, s = 4,800$ obtained by Rogers and Harpending (1992) is shown in Figure 2.7, where the solution to the system of differential equations (2.66,2.77) was used. Also, the expected squared standard linkage deviation in a population, the effective size of which is the harmonic mean of the individual effective sizes over the time period of the growth, or 16,523, is shown, where the solution (2.56) was used. It can be seen that that the expectation of the squared standard linkage deviation for the current entire human population is significantly smaller than that in a constant size population whose effective size is the harmonic mean of the individual effective sizes over the time period of the growth. The squared standard linkage deviation given by the asymptotic formula (2.74) is also shown in Figure 2.7. It can be seen that the time period of the growth is too short to use the asymptotic formula.

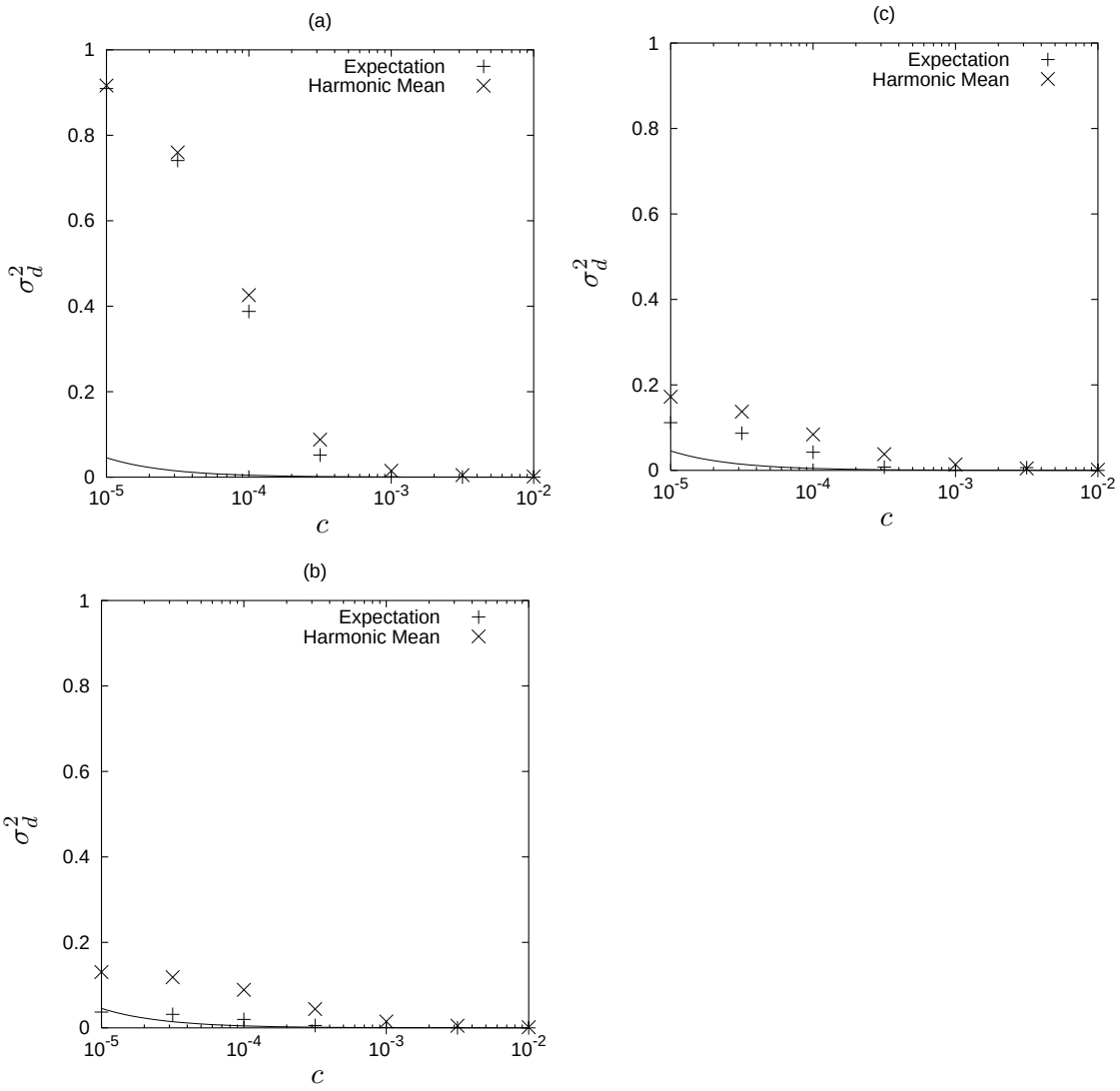

FIGURE 2.7. The expected squared standard linkage deviations in the current whole human population. The size growth model is based on the estimates given by Rogers and Harpending (1992). Those in a constant size population the effective size of which is the harmonic mean of the individual effective sizes over the time period of the growth is also shown (×). (a) the admixture model; (b) the single mutation model; (c) the recurrent mutation model. The lines show values given by the asymptotic formula.

2.8. Appendix. Derivation of equations for the moments

The covariance matrix of $\boldsymbol{x}$ is

$$\boldsymbol{\sigma}^T\boldsymbol{\sigma} = \begin{pmatrix} x(1-x) & z & z(1-2x) \\ z & y(1-y) & z(1-2y) \\ z(1-2x) & z(1-2y) & xy(1-x)(1-y)+z(1-2x)(1-2y)-z^2 \end{pmatrix}. \tag{2.91}$$

By applying the Cholesky decomposition to the matrix, we have a matrix $\boldsymbol{\sigma}$, which is a lower triangular form ($(\boldsymbol{\sigma})_{ij} = 0, i < j, i, j = 1, 2, 3$):

$$\begin{aligned}
(\boldsymbol{\sigma})_{11} &= \sqrt{x(1-x)}, \qquad (\boldsymbol{\sigma})_{21} = z/\sqrt{x(1-x)}, \qquad (\boldsymbol{\sigma})_{31} = z(1-2x)/\sqrt{x(1-x)}, \\
(\boldsymbol{\sigma})_{22} &= \sqrt{y(1-y) - z^2/\{x(1-x)\}}, \\
(\boldsymbol{\sigma})_{32} &= \left[z(1-2y) - z^2(1-2x)/\{x(1-x)\}\right]/\boldsymbol{\sigma}_{22}, \\
(\boldsymbol{\sigma})_{33} &= \sqrt{xy(1-x)(1-y) + z(1-2x)(1-2x) - z^2 - (\boldsymbol{\sigma})_{31}^2 - (\boldsymbol{\sigma})_{32}^2}.
\end{aligned}$$

Applying the Itô formula to a function $x(1-x)y(1-y)$ (Øksendal, 2000), we have

$$\begin{aligned}
d\{x(1-x)y(1-y)\} &= (1-2x)y(1-y)dx + (1-2y)x(1-x)dy - y(1-y)dx^2 \\
&\quad -x(1-x)dy^2 + (1-2x)(1-2y)dxdy \\
&= (1-2x)y(1-y)\boldsymbol{\sigma}_{11}dB_1 + x(1-x)(1-2y)(\boldsymbol{\sigma}_{21} + \boldsymbol{\sigma}_{22})dB_2 \\
&\quad -2x(1-x)y(1-y)d\tau + z(1-2x)(1-2y)dt. \qquad (2.92)
\end{aligned}$$

The Itô integral is

$$\begin{aligned}
x(1-x)y(1-y) &= \int_0^t (1-2x)y(1-y)\boldsymbol{\sigma}_{11}dB_1 + \int_0^t x(1-x)(1-2y)(\boldsymbol{\sigma}_{21} + \boldsymbol{\sigma}_{22})dB_2 \\
&\quad -2\int_0^t x(1-x)y(1-y)d\zeta + \int_0^t z(1-2x)(1-2y)d\zeta, \qquad (2.93)
\end{aligned}$$

and we have

$$\mathbb{E}[x(1-x)y(1-y)] = -2\int_0^t \mathbb{E}[x(1-x)y(1-y)]d\zeta + \int_0^t \mathbb{E}[z(1-2x)(1-2y)]d\zeta. \qquad (2.94)$$

Then,

$$\frac{d\mathbb{E}[X(1-X)Y(1-Y)]}{dt} = -2\mathbb{E}[x(1-x)y(1-y)] + \mathbb{E}[z(1-2x)(1-2y)]. \qquad (2.95)$$

Applying the Itô formula to functions $Z(1-2X)(1-2Y)$ and Z^2 in the same manner, we obtain a system of differential equations (2.52). The method presented here is equivalent to the method to derive equation for the moments which was presented by Ohta and Kimura (1969b).

CHAPTER 3

Concerted Evolution of Duplicated Genes

3.1. Introduction

The evolutionary rate of a gene is defined as the rate of nucleotide substitutions in a certain time period, say, per year or per generation (Zuckerkandl and Pauling, 1965; Jukes and Canter, 1969). The rate is given by the product of the mutation rate and the fixation probability. Under neutrality, it is well known that, in a random mating diploid population with an effective size of N, the neutral evolutionary rate is identical to the mutation rate. When genic selection operates in the locus, the fixation probability of a mutant is

$$u(p) = \frac{1 - e^{-4Nsp}}{1 - e^{-4Ns}}, \tag{3.1}$$

where p is the initial frequency of the mutant and the selective advantage of a mutant allele over a wildtype allele is s (Kimura, 1957). Fixations of mutations also occur in genes that belong to a multigene family. It is possible that a mutation spreads over all member genes of a multigene family when they undergo concerted evolution, a phenomenon that the members evolve in a concerted manner by exchanging their DNA sequences (Ohta, 1980; Dover, 1982). A typical observation of concerted evolution is that the nucleotide sequence diversity between the copy members in the family is very low because of frequent exchanges of genetic material, while there is substantial divergence from the orthologous family in other species. This means that genetic variation among the family can migrate between different copy members, and a certain allele eventually becomes fixed in the species. The accumulation of such fixations results in the divergence between species.

In this chapter, the rate of nucleotide substitutions in duplicated genes or a small multigene family, that are currently undergoing concerted evolution by gene conversion is studied. Gene conversion between copy members should be the major mechanism to cause concerted evolution of small multigene families (Ohta, 1983a), while both gene conversion and unequal crossing over should be working simultaneously in middle- and large-size families (Hillis et al., 1991). The fixation of a mutant which appears in one of the member in a multigene family is that all locus of all individuals of the population are occupied by

the mutant allele at all loci. Theoretical aspects of genetic variation within a multigene family have been extensively studied by Innan (2002, 2003a) and Teshima and Innan (2004). A directional selection model, in which selection works on the copy number of the mutant in a diploid, is assumed. An analytic expression of the fixation probability in terms of a two-locus diffusion model, which was firstly obtained by Mano and Innan (2008), is presented. When no dominance exists between the copy number of the mutant in a diploid, the formula for the fixation probability can be extended to the n-locus model (Mano and Innan, 2008). With the formula of the fixation probabilities, effects of selection on the rate of molecular evolution in a multigene family under concerted evolution are discussed.

It is known that GC-rich regions include many genes in mammalian genomes (Durret et al., 1995). This suggests that the distribution of GC content could have some functional relevance, raising the issue of its origin and evolution. A possible evolutionary force that might explain the pattern is biased gene conversion. Since biased mismatch DNA repair toward GC has been observed experimentally (Brown and Jiricny, 1987), gene conversion could favor particular alleles over others, or GC over AT base pairs. If biased gene conversion were major determinant of GC content evolution, one would expect sequences undergoing frequent gene conversion to become GC-rich. In fact, among genes undergoing concerted evolution in mammals, ribosomal operons, transfer RNAs, and histones are all GC-rich, consistent with the prediction (Galtier, et al. 2001).

In this chapter, a model of biased gene conversion, in which gene conversion favor an allelic type over the other type, is studied. An analytic expression of the fixation probability in terms of an n-locus diffusion model is presented. With the formula of the fixation probability, effects of biased gene conversion on the rate of molecular evolution in a multigene family under concerted evolution are discussed.

3.2. A two-locus diffusion model with selection and gene conversion

Let us assume two loci in which pair of alleles A, a are segregating, and let the initial frequencies of gametes AA, Aa, aA, aa be respectively g_1, g_2, g_3 and $1-(g_1+g_2+g_3)$, and let the frequencies at time t be respectively x_1, x_2, x_3 and $1-x_1-x_2-x_3$. Let the initial frequencies of allele A at the first and the second locus be respectively p and q, and the frequency of them at time t be respectively x and y. Let $D = g_1(1-g_1-g_2-g_3)-g_2g_3$ be the initial value of the linkage disequilibrium coefficient and $z = x_1(1-x_1-x_2-x_3)-x_2x_3$ be the value of the linkage disequilibrium coefficient at time t. Assume zygotes consist of random union of two gametes. Let relative fitness of zygotes which have i copies of allele A

to be $1+s_i,\ i=0,1,2,3,4$, where $s_0=0$. Let the recombination fraction between two loci be $r\ (>0)$ and the conversion rate between the two loci be $c\ (>0)$. A diffusion time scaling is to measure time in units of $4N$ generations and let $4N\to\infty$, while $\sigma_i=4Ns_i, \rho=4Nr$ and $\gamma=4Nc$ are held constant. The Wright-Fisher model converges to a diffusion process $\{x(t),y(t),z(t);t\geq 0\}$ in H (see Chapter 1), which is governed by a generator

$$
\begin{aligned}
L &= x(1-x)\frac{\partial^2}{\partial x^2}+y(1-y)\frac{\partial^2}{\partial y^2}+\{xy(1-x)(1-y)+z(1-2x)(1-2y)-z^2\}\frac{\partial^2}{\partial z^2}\\
&\quad +2z\frac{\partial^2}{\partial x\partial y}+2z(1-2x)\frac{\partial^2}{\partial x\partial z}+2z(1-2y)\frac{\partial^2}{\partial y\partial z}+\{-\gamma(x-y)+\sigma_x\}\frac{\partial}{\partial x}\\
&\quad +\{\gamma(x-y)+\sigma_y\}\frac{\partial}{\partial y}+\{\gamma x(1-x)+\gamma y(1-y)-(2+2\gamma+\rho)z+\sigma_z\}\frac{\partial}{\partial z}, \qquad (3.2)
\end{aligned}
$$

where

$$
\begin{aligned}
\sigma_x &= \sigma_1x_4\{x_2-2(x_1+x_2)(x_2+x_3)\}\\
&\quad +\sigma_2[x_1x_4\{1-2(x_1+x_2)\}+\{x_2-(x_1+x_2)(x_2+x_3)\}(x_2+x_3)]\\
&\quad +\sigma_3x_1\{2x_2+x_3-2(x_1+x_2)(x_2+x_3)\}+\sigma_4x_1^2(1-x_1-x_2),\\
\sigma_y &= \sigma_1x_4\{x_3-2(x_3+x_1)(x_2+x_3)\}\\
&\quad +\sigma_2[x_1x_4\{1-2(x_3+x_1)\}+\{x_3-(x_3+x_1)(x_2+x_3)\}(x_2+x_3)\}\\
&\quad +\sigma_3x_1\{x_2+2x_3-2(x_3+x_1)(x_2+x_3)\}+\sigma_4x_1^2(1-x_3-x_1),\\
\sigma_z &= \sigma_1x_4\{-3x_1x_2-2x_2x_3-3x_3x_1+4(x_1+x_2)(x_2+x_3)(x_3+x_1)\}\\
&\quad +\sigma_2[x_1x_4\{1-4x_1-x_2-x_3+4(x_1+x_2)(x_1+x_3)\}\\
&\quad +2(x_2+x_3)(-x_1x_2-x_2x_3-x_3x_1+(x_1+x_2)(x_2+x_3)(x_3+x_1)\}\\
&\quad +\sigma_3x_1\{-(x_2+x_3+5x_1)(x_2+x_3)-2x_2x_3-4(x_1+x_2)(x_2+x_3)(x_3+x_1)\}\\
&\quad +\sigma_4x_1^2\{1-3x_1-x_2-x_3-2(x_1+x_2)(x_1+x_3)\}.
\end{aligned}
$$

The generator is an extension of that obtained by Innan (2002). Let us count degree of terms by sum of degree in x, y and twice of z. For example, degree of xyz is counted by $1+1+2=4$. Then, in general, σ_x and σ_y involve terms up to fifth degree. For computational simplicity, we put the condition that σ_x and σ_y involves up to third degree terms: $\sigma_3=-3\sigma_1+3\sigma_2$ and $\sigma_4=-8\sigma_1+6\sigma_2$. Then, we can reparametrize the selection coefficients by two parameters σ and h:

$$
\sigma_1=\left(3h-\frac{1}{2}\right)\sigma,\qquad \sigma_2=4h\sigma,\qquad \sigma_3=\left(3h+\frac{3}{2}\right)\sigma,\qquad \sigma_4=4\sigma. \qquad (3.3)
$$

Especially, $h = 1/2$ corresponds to the additive effect model. The coefficients are simplified and

$$\begin{aligned}
\sigma_x &= \sigma\left[(1-2h)\{x(1-x)(x+2y)+yz\}+\left(3h-\frac{1}{2}\right)x(1-x)+\left(h+\frac{1}{2}\right)z\right], \\
\sigma_y &= \sigma\left[(1-2h)\{y(1-y)(y+2x)+yz\}+\left(3h-\frac{1}{2}\right)y(1-y)+\left(h+\frac{1}{2}\right)z\right], \\
\sigma_z &= \sigma[(2h-1)\{z^2-xy(1-x)(1-y)+2(x+y)(x+y-3)z\} \\
&\quad -4(1-h)(x+y)z+4hz]. \qquad (3.4)
\end{aligned}$$

3.3. Fixation probability with selection

Both of the two loci destined to be fixed by either of the allele A or a. Denote the fixation probability of the allele A be $u(p,q,D)$. $u(p,q,D)$ satisfies the following partial differential equation

$$Lu(p,q,D) = 0. \qquad (3.5)$$

To address the boundary condition, it is convenient to consider the diffusion process $\{x_1(t), x_2(t), x_3(t); t \geq 0\}$ in K. The fixation probability $u(g_1,g_2,g_3)$ satisfies the elliptic partial differential equation

$$\sum_{i,j=1}^{3} g_i(\delta_{i,j}-g_j)\frac{\partial^2 u}{\partial g_i^2}+\sum_{i=1}^{3} v_i(g_1,g_2,g_3)\frac{\partial u}{\partial g_i}=0. \qquad (3.6)$$

The generator degenerates over ∂K. It is obvious that $(x_1,x_2,x_3) = (1,0,0),(0,0,0)$ are the exit boundaries, and we may put the Dirichlet type conditions $u(1,0,0) = 1, u(0,0,0) = 0$. However, appropriate conditions for other portion of the boundary

$$S = \partial K - \{(1,0,0),(0,0,0)\} \qquad (3.7)$$

is not clear. In contrast to the classification of boundary conditions for one dimensional processes (Feller, 1952), boundary conditions for higher dimensional processes is not thoroughly studied. It is known that we do not always have to impose conditions at all portion of a boundary. But the existing general theory is not applicable here, since it depends on C^2-differentiability of the boundary (Oleinik and Radkevic, 1973), while ∂K is not differentiable at the edges. So, we employ biological intuition to put the condition on S. Since there should be finite probability of fixation when a population evolves from an arbitrary point in S, we may put a condition

$$u(g_1,g_2,g_3)|_s \in (0,1). \qquad (3.8)$$

For example, consider a point $(\epsilon, c_2, c_3) \in K$ near to the surface

$$S_1 : x_1 = 0, 0 < x_2 < x_2 + x_3 < 1 \tag{3.9}$$

in S with small ϵ (> 0). We may expand u as a Laurent series in ϵ

$$u(\epsilon, c_2, c_3) = \sum_{i=-\infty}^{\infty} f_i(c_2, c_3)\epsilon^i, \tag{3.10}$$

where $f_i(c_2, c_3)$ is a function of c_2 and c_3. To keep $u(\epsilon, c_2, c_3)$ be finite with $\epsilon \to 0$, we have to set $f_i = 0, i < 0$. Subsequently, we have

$$u(g_1, g_2, g_3)|_{S_1} = \lim_{\epsilon \to 0} u(\epsilon, c_2, c_3) = f_0(c_2, c_3) \in (0, 1). \tag{3.11}$$

This kind of argument on boundary condition has been used in population genetics literature to discuss fixation time of an allele (for example, Kimura and Ohta (1969); Kimura and King (1979)).

The additive effect model ($h = 1/2$) has special interest. For the case, it is straightforward to confirm that

$$u(p, q, D) = \frac{1 - e^{-2\sigma\bar{p}}}{1 - e^{-2\sigma}}, \qquad \bar{p} = \frac{p + q}{2} \tag{3.12}$$

is the solution. Actuary, it satisfies (3.5), $u(1, 1, 0) = 1, u(0, 0, 0) = 0$. Also, it is straightforward to check that it satisfies (3.8) for each portion of S. For example,

$$u(x_1, x_2, x_3)|_{S_1} = \frac{1 - e^{-\sigma(x_2+x_3)}}{1 - e^{-2\sigma}} \in (0, 1). \tag{3.13}$$

The two-locus fixation probability (3.12) is identical to the single-locus fixation probability (3.1) with replacing σ by 2σ and p by $\bar{p}$. The two-locus model discussed here is substantially complicated than the single-locus model, nevertheless, the fixation probabilities of them are almost identical. The two-locus fixation probability is independent of D, ρ and γ. Walsh (1985) has obtained an analytical expression for the fixation probability with the assumption that gene conversion rate is low and under linkage equilibrium among loci (Equation 8 in Walsh (1985)). The expression (3.12) reduces to Walsh's result in the limit. In addition, (3.12) is consistent with the result of Nagylaki and Petes (1982), who obtained the fixation probability with the assumption of infinite population size and under neutrality.

Furthermore, Mano and Innan (2008) showed that (3.12) can be extended to models where more than two loci are involved. By simulations, they demonstrated that

$$u(p, q, D) = \frac{1 - e^{-n\sigma\bar{p}}}{1 - e^{-n\sigma}}, \tag{3.14}$$

where $\bar{p}$ is the arithmetic mean of the initial allele frequencies of A in each locus, holds for $n = 3, 4, 5, 6$ loci. It is worth pointing out that the expression (3.14) also appear in a problem with population structure and migration. Maruyama (1972) obtained the identical expression as the fixation probability of a mutant under genic selection, where the mutant spreads over the whole population which consists of n subpopulations whose effective sizes are N. The remarkable property is that the expression does not depend on properties of migration pattern, as long as the population sizes of each subpopulation are maintained. It is straightforward to show that n-unlinked locus version of the generator (3.2) is identical to the generator of the process of a n-island model with the migration rates between any of subpopulations are γ. Thus, Maruyama's solution should be the same as that for our n-locus model, at least when $\rho = \infty$.

It seems difficult to obtain the exact solution of (3.5) in general case with $h \neq 1/2$. But it is possible to obtain a perturbative expansion of it. Denote the moments

$$\mu_{l,m,n}(t) = \mathbb{E}[x^l y^m z^n], \qquad l, m, n = 0, 1, ... \tag{3.15}$$

Because $(x, y, z) = (1, 1, 0), (0, 0, 0)$ are the exit boundaries, we should have

$$\lim_{t\to\infty} \mu_{l,m,0}(t) = \lim_{t\to\infty} \mathbb{E}[x^l y^m] = \mathbb{P}[x(\infty) = y(\infty) = 1] = u(p, q, D), \qquad l + m = 1, 2, ... \tag{3.16}$$

Thus, we can obtain $u(p, q, D)$ as a limit of a moment $\mu_{1,0,0}(t)$. Assume σ is not so large such that the solution can be well expressed by a perturbative series in σ with few terms

$$\mu_{l,m,n}(t) = \mu^{(0)}_{l,m,n}(t) + \sigma\mu^{(1)}_{l,m,n}(t) + \sigma^2\mu^{(2)}_{l,m,n}(t) + \cdots \tag{3.17}$$

The initial condition is $\mu^{(0)}_{l,m,n}(0) = p^l q^m D^n$ and $\mu^{(i)}_{l,m,n}(0) = 0, i = 1, 2, ...$ The diffusion process governed by the generator (3.2) is represented by a system of stochastic differential equations (See Chapter 1)

$$d\boldsymbol{x} = \sigma d\boldsymbol{B} + \boldsymbol{v}dt, \tag{3.18}$$

where

$$\boldsymbol{v} = (-\gamma(p - q) + \sigma_x, \gamma(p - q) + \sigma_y, \gamma\{x(1 - x) + y(1 - y)\} - z(2 + 2\gamma + \rho) + \sigma_z)'. \tag{3.19}$$

By using the Itô formula, we obtain differential equations for the moments (see Chapter 1). For example, applying the Itô formula to a function x, we have

$$\begin{aligned} \frac{d\mu_{1,0,0}}{dt} &= -\gamma(\mu_{1,0,0} - \mu_{0,1,0}) + \sigma\left\{\left(3h - \frac{1}{2}\right)\mu_{1,0,0} + \left(h + \frac{1}{2}\right)\mu_{0,0,1} + \left(\frac{3}{2} - 5h\right)\mu_{2,0,0}\right. \\ &\quad \left. +(1 - 2h)(2\mu_{1,1,0} - \mu_{3,0,0} + \mu_{0,1,1} - 2\mu_{2,1,0})\right\}. \end{aligned} \tag{3.20}$$

Substituting (3.17) into the equation (3.20), we have ordinary differential equations for each order of the perturbation. For the zeroth order in σ, we have

$$\frac{d\mu^{(0)}_{1,0,0}}{dt} = -\gamma(\mu^{(0)}_{1,0,0} - \mu^{(0)}_{0,1,0}), \tag{3.21}$$

and for the higer orders, we have

$$\begin{aligned}\frac{d\mu^{(i)}_{1,0,0}}{dt} &= -\gamma(\mu^{(i)}_{1,0,0} - \mu^{(i)}_{0,1,0}) + \left\{ \left(3h - \frac{1}{2}\right)\mu^{(i-1)}_{1,0,0} + \left(h + \frac{1}{2}\right)\mu^{(i-1)}_{0,0,1} + \left(\frac{3}{2} - 5h\right)\mu^{(i-1)}_{2,0,0} \right. \\ &\left. +(1-2h)(2\mu^{(i-1)}_{1,1,0} - \mu^{(i-1)}_{3,0,0} + \mu^{(i-1)}_{0,1,1} - 2\mu^{(i-1)}_{2,1,0}) \right\}, \qquad i = 1, 2, ... \end{aligned} \tag{3.22}$$

Let us obtain $\mu^{(0)}_{1,0,0}$ and $\mu^{(1)}_{1,0,0}$. First, $\mu^{(1)}_{1,0,0}$ and $\mu^{(1)}_{0,1,0}$ satisfy

$$\begin{aligned}\frac{d\mu^{(1)}_{1,0,0}}{dt} &= -\gamma(\mu^{(1)}_{1,0,0} - \mu^{(1)}_{0,1,0}) + \left(3h - \frac{1}{2}\right)\mu^{(0)}_{1,0,0} + \left(h + \frac{1}{2}\right)\mu^{(0)}_{0,0,1} + \left(\frac{3}{2} - 5h\right)\mu^{(0)}_{2,0,0} \\ &+(1-2h)(2\mu^{(0)}_{1,1,0} - \mu^{(0)}_{3,0,0} + \mu^{(0)}_{0,1,1} - 2\mu^{(0)}_{2,1,0}) \end{aligned} \tag{3.23}$$

$$\begin{aligned}\frac{d\mu^{(1)}_{0,1,0}}{dt} &= \gamma(\mu^{(1)}_{1,0,0} - \mu^{(1)}_{0,1,0}) + \left(3h - \frac{1}{2}\right)\mu^{(0)}_{0,1,0} + \left(h + \frac{1}{2}\right)\mu^{(0)}_{0,0,1} + \left(\frac{3}{2} - 5h\right)\mu^{(0)}_{0,2,0} \\ &+(1-2h)(2\mu^{(0)}_{1,1,0} - \mu^{(0)}_{0,3,0} + \mu^{(0)}_{1,0,1} - 2\mu^{(0)}_{1,2,0}), \end{aligned} \tag{3.24}$$

respectively. The Laplace transform of (3.23,3.24) are

$$\begin{aligned}\lambda\nu^{(1)}_{1,0,0} &= -\gamma(\nu^{(1)}_{1,0,0} - \nu^{(1)}_{0,1,0}) + \left(3h - \frac{1}{2}\right)\nu^{(0)}_{1,0,0} + \left(h + \frac{1}{2}\right)\nu^{(0)}_{0,0,1} + \left(\frac{3}{2} - 5h\right)\nu^{(0)}_{2,0,0} \\ &+(1-2h)(2\nu^{(0)}_{1,1,0} - \nu^{(0)}_{3,0,0} + \nu^{(0)}_{0,1,1} - 2\nu^{(0)}_{2,1,0}), \end{aligned} \tag{3.25}$$

$$\begin{aligned}\lambda\nu^{(1)}_{0,1,0} &= \gamma(\nu^{(1)}_{1,0,0} - \nu^{(1)}_{0,1,0}) + \left(3h - \frac{1}{2}\right)\nu^{(0)}_{0,1,0} + \left(h + \frac{1}{2}\right)\nu^{(0)}_{0,0,1} + \left(\frac{3}{2} - 5h\right)\nu^{(0)}_{0,2,0} \\ &+(1-2h)(2\nu^{(0)}_{1,1,0} - \nu^{(0)}_{0,3,0} + \nu^{(0)}_{1,0,1} - 2\nu^{(0)}_{1,2,0}). \end{aligned} \tag{3.26}$$

The expression involves moments up to the third degree. The first degree moments are closed, and we have

$$\begin{pmatrix} \lambda+\gamma & -\gamma \\ -\gamma & \lambda+\gamma \end{pmatrix} \begin{pmatrix} \nu^{(0)}_{1,0,0} \\ \nu^{(0)}_{0,1,0} \end{pmatrix} = \begin{pmatrix} p \\ q \end{pmatrix}. \tag{3.27}$$

For the second degree moments are closed, we have

$$(A_2 + \lambda E) \begin{pmatrix} \nu^{(0)}_{2,0,0} \\ \nu^{(0)}_{0,2,0} \\ \nu^{(0)}_{1,1,0} \\ \nu^{(0)}_{0,0,1} \end{pmatrix} = \begin{pmatrix} 2\nu^{(0)}_{1,0,0} + p^2 \\ 2\nu^{(0)}_{0,1,0} + q^2 \\ pq \\ \gamma(\nu^{(0)}_{1,0,0} + \nu^{(0)}_{0,1,0}) + D \end{pmatrix}, \tag{3.28}$$

where E is the unit matrix, and

$$A_2 = \begin{pmatrix} 2(1+\gamma) & 0 & -2\gamma & 0 \\ 0 & 2(1+\gamma) & -2\gamma & 0 \\ -\gamma & -\gamma & 2\gamma & -2 \\ \gamma & \gamma & 0 & 2+2\gamma+\rho \end{pmatrix}.$$

For the third degree moments are closed, and we have

$$(A_3+\lambda E)\begin{pmatrix} \nu^{(0)}_{3,0,0} \\ \nu^{(0)}_{0,3,0} \\ \nu^{(0)}_{2,1,0} \\ \nu^{(0)}_{1,2,0} \\ \nu^{(0)}_{1,0,1} \\ \nu^{(0)}_{0,1,1} \end{pmatrix} = \begin{pmatrix} 6\nu^{(0)}_{2,0,0}+p^3 \\ 6\nu^{(0)}_{0,2,0}+q^3 \\ 2\nu^{(0)}_{1,1,0}+p^2q \\ 2\nu^{(0)}_{1,1,0}+pq^2 \\ 2\nu^{(0)}_{0,0,1}+\gamma(\nu_{2,0,0}+\nu_{1,1,0})+pD \\ 2\nu^{(0)}_{0,0,1}+\gamma(\nu_{0,2,0}+\nu_{1,1,0})+qD \end{pmatrix}, \tag{3.29}$$

where

$$A_3 = \begin{pmatrix} 6+3\gamma & 0 & -3\gamma & 0 & 0 & 0 \\ 0 & 6+3\gamma & 0 & -3\gamma & 0 & 0 \\ -\gamma & 0 & 2+3\gamma & -2\gamma & -4 & 0 \\ 0 & -\gamma & -2\gamma & 2+3\gamma & 0 & -\rho \\ \gamma & 0 & 0 & \gamma & 6+3\gamma+\rho & -\gamma \\ 0 & \gamma & \gamma & 0 & -\gamma & 6+3\gamma+\rho \end{pmatrix}.$$

It is straightforward to solve (3.27), and we have

$$\nu^{(0)}_{1,0,0} = \frac{\bar{p}}{\lambda} + \frac{p-q}{2(\lambda+2\gamma)}. \tag{3.30}$$

By substituting solution of (3.27, 3.28,3.29) into (3.25, 3.26), we have

$$\nu^{(1)}_{1,0,0} = \frac{\bar{p}(1-\bar{p}) + \left(h-\frac{1}{2}\right)u^{(1)}_d}{\lambda} + \sum_{i=1}^{10}\frac{a_i}{\lambda-\lambda_i}, \tag{3.31}$$

where

$$u^{(1)}_d = (1-2\bar{p})\left[\bar{p}(1-\bar{p})\left\{\frac{2}{3}+\frac{4\rho\gamma}{(1+\gamma)(6+4\gamma+\rho)}\right\} - \frac{(3+2\gamma)D}{(1+\gamma)(6+4\gamma+\rho)} - \frac{pq-\bar{p}}{2(1+\gamma)}\right],$$

and λ_i $(<0), i = 1, 2, ..., 10$ are eigenvalues of A_2 and A_3, and the coefficients a_i depend on p, q, D, h, ρ, γ. Especially, eigenvalues of A_2 are $\lambda_1 = -2(1+\gamma)$ and the three roots of a cubic equation

$$\xi^3 + (\rho-2)\xi^2 - 2(\rho+2\gamma^2)\xi + 4\gamma^2(2-\rho) = 0, \tag{3.32}$$

where $\xi = \lambda + 2(1+\gamma)$. This is identical to the equation which was obtained by Ohta (1983b), where x in her Eq. 4 is replaced by $4N\lambda + 1$ in our notation. Let $\lambda_2, \lambda_3, \lambda_4$ be the three roots and let λ_2 be the largest. We have

$$\lambda_2 = -2\gamma - \frac{4+\rho}{3} + 2\sqrt{\frac{\zeta}{3}}\cos\frac{\varphi}{3}, \qquad \cos\varphi = -\frac{\eta}{2}\left(\frac{3}{\zeta}\right)^{3/2}, \qquad 0 \le \varphi \le \pi, \tag{3.33}$$

where

$$\zeta = \frac{(\rho+1)^2}{3} + 1 + 4\gamma^2, \qquad \eta = \frac{2}{27}(\rho-2)\{(\rho+1)(\rho+4) - 36\gamma^2\}.$$

Ohta (1983b) discussed fixation time of a neutral mutant by computing decay rate of genetic identity. In our analysis of fixation probability, fixation time of a neutral mutant can be discussed by computing decay rate of the moments. Up to the second degree moments, the largest eigenvalue $\max\{-2\gamma, \lambda_2\}$ gives asymptotic decay rate after large number of generations. Since $\max\{-2\gamma, \lambda_2\} = \lambda_2$ (Because left hand side of the cubic equation (3.32) is $-4\gamma^2\rho \le 0$ at $\xi = 2$, we have $\lambda_2 + 2(1+\gamma) \ge 2$), the decay rate is λ_2, which is in accordance with Ohta's estimate.

Applying the inverse Laplace transformation to (3.30,3.31), we obtain

$$\mu^{(0)}_{1,0,0} = \bar{p} + \frac{p-q}{2}e^{-2\gamma t}, \tag{3.34}$$

$$\mu^{(1)}_{1,0,0} = \bar{p}(1-\bar{p}) + \left(h - \frac{1}{2}\right)u^{(1)}_d + \tilde{\mu}^{(1)}_{0,1,0}(t), \tag{3.35}$$

where $\tilde{\mu}^{(1)}_{0,1,0}(t)$ represent terms decay as $t \to \infty$. Thus, up to the first order in σ, we have

$$u(p,q,D) = \bar{p} + \sigma\left\{\bar{p}(1-\bar{p}) + \left(h - \frac{1}{2}\right)u^{(1)}_d\right\} + O(\sigma^2). \tag{3.36}$$

For the case without dominance $h = 1/2$, it is straightforward to check that (3.36) is consistent with the exact expression (3.12). The correspondence between the two-locus fixation probability and the single-locus fixation probability does not hold in general. The single-locus fixation probability is (Kimura, 1957)

$$\begin{aligned} u(p) &= \int_0^p e^{-\sigma z[1+(2h-1)(1-z)]}dz \Big/ \int_0^1 e^{-\sigma z[1+(2h-1)(1-z)]}dz \\ &= p + \frac{\sigma}{2}\left\{p(1-p) + \frac{2}{3}\left(h - \frac{1}{2}\right)p(1-p)(1-2p)\right\} + O(\sigma^2), \end{aligned} \tag{3.37}$$

where relative fitness of zygotes for AA, Aa and aa are $1+2s, 1+2hs$ and $1+s$, respectively. (3.36) and (3.37) are not identical when we replace σ by 2σ and p by $\bar{p}$. Nevertheless, for $\gamma \to \infty$, up to the first order in σ, the two-locus fixation probability (3.36) and (3.37) are identical with replacing σ by 2σ and p by $\bar{p}$.

3.4. An n-locus diffusion model with biased gene conversion

Assume gene conversion occurs among the n unlinked duplicated loci with rates per locus par a single generation of $2\alpha c$ for conversion from a gene A to a gene a gene and with $2(1-\alpha)c$ for conversion of the reverse direction, where $1/2 \leq \alpha \leq 1$. Assume the bias is not strong such that $\alpha = 1/2 + \epsilon$ with small ϵ (> 0). Let the initial frequencies of the allele A in the i-th locus be $p_i, i = 1, 2, ..., n$, and let the frequencies at time t be respectively $x_i, i = 1, 2, ..., n$. Denote type of gametes by a binary number, which has 1 in loci with the allele A and has 0 in loci with the allele a. Let frequency of a gamete whose type is $\boldsymbol{g}$ be $f(\boldsymbol{g})$, where $g_i = 1, 0$ when the i-th locus of the gamete $\boldsymbol{g}$ is occupied by a gene A and a, respectively. Since these loci are unlinked, we have

$$f(\boldsymbol{g}) = \prod_{\{i:g_i=1\}} x_i \prod_{\{i:g_i=0\}} (1-x_i). \tag{3.38}$$

The increments of the allele frequencies in a single generation is, for $i = 1, 2, ..., n$,

$$\begin{aligned} \delta x_i &= 2\alpha c \sum_{j=1}^{n-1} j f(g_i = 0, |\boldsymbol{g}| = j) - 2(1-\alpha)c \sum_{j=1}^{n-1} j f(g_i = 1, |\boldsymbol{g}| = n-j) \\ &= c\{\sum_{j\neq i} x_j - (n-1)x_i\} + 2c\epsilon\{(1-2x_i)\sum_{j\neq i} x_j + (n-1)x_i\}. \end{aligned} \tag{3.39}$$

A diffusion time scaling is to measure time in units of $4N$ generations and let $4N \to \infty$, while $\gamma = 4Nc$ is held constant. The Wright-Fisher model converges to a n-dimensional diffusion process $\{x_i(t), i = 1, 2, ..., n; t \geq 0\}$ in a n-dimensional hyper cube $[0, 1]^n$ with a generator

$$L = L_0 + \epsilon L_1, \tag{3.40}$$

where

$$\begin{aligned} L_0 &= \sum_{i=1}^{n} x_i(1-x_i)\frac{\partial^2}{\partial x_i^2} + \gamma \sum_{i=1}^{n} \{\sum_{j\neq i} x_j - (n-1)x_i\}\frac{\partial}{\partial x}, \\ L_1 &= 2\gamma \sum_{i=1}^{n} \{(1-2x_i)\sum_{j\neq i} x_j + (n-1)x_i\}\frac{\partial}{\partial x}. \end{aligned}$$

3.5. Fixation probability with biased gene conversion

All locus destined to be fixed by either of the allele A or a. Denote the fixation probability of the allele A be $u(p_1, p_2, ..., p_n)$. $u(p_1, p_2, ..., p_n)$ satisfies a partial differential equation

$$Lu(p_1, p_2, ..., p_n) = 0 \tag{3.41}$$

with a boundary condition (see Section 2.3)

$$u(0,0,...,0)=0, \qquad u(1,1,...,1)=1, \qquad u|_{[0,1]^n-\{(0,0,...,0),(1,1,...,1)\}}=\text{finite}. \tag{3.42}$$

Suppose ϵ is not so large such that the solution can be well expressed by a perturbative series in ϵ with few terms

$$u(p_1,p_2,...,p_n)=u^{(0)}+\epsilon u^{(1)}(p_1,p_2,...,p_n)+\epsilon^2 u^{(2)}(p_1,p_2,...,p_n)+\cdots \tag{3.43}$$

At the zeroth order in ϵ, we have a partial differential equation

$$L_0 u^{(0)}=0 \tag{3.44}$$

with the boundary condition

$$u^{(0)}(0,0,...,0)=0, \qquad u^{(0)}(1,1,...,1)=1, \qquad u^{(0)}|_{[0,1]^n-\{(0,0,...,0),(1,1,...,1)\}}=\text{finite}. \tag{3.45}$$

It is straightforward to obtain the solution

$$u^{(0)}(p_1,p_2,...,p_n)=\bar{p}, \qquad \bar{p}=\sum_{i=1}^{n}\frac{p_i}{n}. \tag{3.46}$$

At the first order in ϵ, we have a partial differential equation

$$L_0 u^{(1)}=-L_1 u^{(0)}=4(n-1)\gamma\bar{p}-\frac{8\gamma}{n}\sum_{i<j}p_i p_j, \tag{3.47}$$

with the boundary condition

$$u^{(1)}(0,0,...,0)=u^{(1)}(1,1,...,1)=0, \qquad u^{(1)}|_{[0,1]^n-\{(0,0,...,0),(1,1,...,1)\}}=\text{finite}. \tag{3.48}$$

It is straightforward to confirm that

$$u^{(1)}(p_1,p_2,...,p_n)=2(n-1)\bar{p}-\frac{4}{n}\sum_{i<j}p_i p_j+2n(n-1)\gamma\bar{p}(1-\bar{p}) \tag{3.49}$$

is the solution. Actually, it satisfies (3.47) with (3.48). Note that $u^{(1)}$ does not vanish as $\gamma\to 0$. The acceleration of the fixation probability does not vanish in the limit of week gene conversion.

It is possible to obtain (3.49) by taking limit of moments. Because $(0,0,...,0)$ and $(1,1,...,1)$ are the exit boundaries, we should have for $i=1,2,...,n$

$$\lim_{t\to\infty}\mathbb{E}[x_i]=\mathbb{P}[x_1(\infty)=x_2(\infty)=\cdots=x_n(\infty)=1]=u(p_1,p_2,...,p_n). \tag{3.50}$$

Let for $c_1,c_2,...c_n=0,1,...$

$$\mu_{c_1c_2...c_n}(t)=\mathbb{E}[\prod_{i=1}^{n}p_i^{c_i}(t)]=\mu^{(0)}_{c_1c_2...c_n}(t)+\epsilon\mu^{(1)}_{c_1c_2...c_n}(t)+\cdots. \tag{3.51}$$

The initial condition is $\mu^{(0)}_{c_1c_2...c_n}(0) = \prod_{i=1}^n p_i^{c_i}$ and $\mu^{(1)}_{c_1c_2...c_n}(0) = 0$. Denote the Laplace transforms of $\mu^{(i)}_{c_1c_2...c_n}(t)$ by $\nu^{(i)}_{c_1c_2...c_n}(\lambda)$. From (3.50), it is clear that

$$u^{(1)}(p_1, p_2, ..., p_n) = \lim_{t\to\infty} \mu^{(1)}_{c_i=1,\{c_j=0:j\neq i\}}(t), \qquad i = 1, 2, ..., n. \tag{3.52}$$

For the zeroth order (in ϵ) moments of the first degree (where degree is counted by degree in $p_i, i = 1, 2, ..., n$), by taking expectations of the stochastic differential equation with (3.40), we have

(3.53)

$$\lambda \begin{pmatrix} \nu^{(0)}_{10...0} \\ \nu^{(0)}_{010...0} \\ . \\ . \\ \nu^{(0)}_{0...1} \end{pmatrix} - \begin{pmatrix} p_1 \\ p_2 \\ . \\ . \\ p_n \end{pmatrix} = \begin{pmatrix} -(n-1)\gamma & \gamma & ... & \gamma \\ \gamma & -(n-1)\gamma & ... & \gamma \\ . & . & ... & . \\ . & . & ... & . \\ \gamma & \gamma & ... & -(n-1)\gamma \end{pmatrix} \begin{pmatrix} \nu^{(0)}_{10...0} \\ \nu^{(0)}_{010...0} \\ . \\ . \\ \nu^{(0)}_{0...1} \end{pmatrix}$$

and the solution is

$$\nu^{(0)}_{c_i=1,\{c_j=0:j\neq i\}}(\lambda) = \frac{\bar{p}}{\lambda} + \frac{p_i - \bar{p}}{\lambda + n\gamma}, \qquad i = 1, 2, ..., n. \tag{3.54}$$

By the inverse Laplace transform, we have

$$\mu^{(0)}_{c_i=1,\{c_j=0:j\neq i\}}(t) = \bar{p} + (p_i - \bar{p})e^{-n\gamma t}, \qquad i = 1, 2, ..., n. \tag{3.55}$$

It is worth to be mentioned that a measure of time to the fixation is given by $n\gamma$. For the second degree moments, by using the Itô formula (see Chapter 1), it is straightforward to obtain

$$\begin{aligned} \nu^{(0)}_{c_i=2,\{c_j=0:j\neq i\}} &= p_i^2 - \{2 + 2(n-1)\gamma\}\nu^{(0)}_{c_i=2,\{c_j=0:j\neq i\}} + 2\gamma \sum_{j\neq i} \nu^{(0)}_{c_i=c_j=1,\{c_k=0:k\neq i,j\}} \\ &\quad +2\nu^{(0)}_{c_i=1,\{c_j=0:j\neq i\}}, \qquad i = 1, 2, ..., n, \end{aligned} \tag{3.56}$$

and for $i \neq j, i, j = 1, 2, ..., n-1$,

$$\begin{aligned} &\nu^{(0)}_{c_i=c_j=1,\{c_k=0:k\neq i,j\}} = -2(n-1)\gamma\nu^{(0)}_{c_i=c_j=1,\{c_k=0:k\neq i,j\}} + \gamma\nu^{(0)}_{c_i=2,\{c_k=0:k\neq i\}} \\ &+\gamma\nu^{(0)}_{c_j=2,\{c_k=0:k\neq j\}} + \gamma \sum_{k\neq i,j} \nu^{(0)}_{c_j=c_k=1,\{c_l=0:l\neq j,k\}} + \gamma \sum_{k\neq i,j} \nu^{(0)}_{c_i=c_k=1,\{c_l=0:l\neq i,k\}}. \end{aligned} \tag{3.57}$$

Substituting (3.54) into (3.56), we have the system of $n(n+1)/2$ equations for the moments (3.56) and (3.57). They are closed and it is possible to be solved. For the first order

moments, we have

$$\lambda\begin{pmatrix}\nu^{(1)}_{10...0}\\ \nu^{(1)}_{010...0}\\ .\\ .\\ \nu^{(1)}_{0...1}\end{pmatrix}=\begin{pmatrix}-(n-1)\gamma & \gamma & ... & \gamma\\ \gamma & -(n-1)\gamma & ... & \gamma\\ . & . & ... & .\\ . & . & ... & .\\ \gamma & \gamma & ... & -(n-1)\gamma\end{pmatrix}\begin{pmatrix}\nu^{(1)}_{10...0}\\ \nu^{(1)}_{010...0}\\ .\\ .\\ \nu^{(1)}_{0...1}\end{pmatrix}$$
$$+2\gamma\begin{pmatrix}(n-1)\nu^{(0)}_{10...0}+\sum_{i\neq 1}\nu^{(0)}_{c_i=1,\{c_j=0:j\neq i\}}\\ (n-1)\nu^{(0)}_{010...0}+\sum_{i\neq 2}\nu^{(0)}_{c_i=1,\{c_j=0:j\neq i\}}\\ .\\ .\\ (n-1)\nu^{(0)}_{0...01}+\sum_{i\neq n}\nu^{(0)}_{c_i=1,\{c_j=0:j\neq i\}}\end{pmatrix}-4\gamma\begin{pmatrix}\sum_{i\neq 1}\nu^{(0)}_{c_1=c_i=1,\{c_j=0:j\neq 1,i\}}\\ \sum_{i\neq 2}\nu^{(0)}_{c_2=c_i=1,\{c_j=0:j\neq 2,i\}}\\ .\\ .\\ \sum_{i\neq n}\nu^{(0)}_{c_n=c_i=1,\{c_j=0:j\neq n,i\}}\end{pmatrix}$$

(3.58)

Substituting (3.54) and solutions for (3.56) and (3.57) into (3.58), we have for $i=1,2,...,n$

(3.59)

$$\nu^{(1)}_{c_i=1,\{c_j=0:j\neq i\}}(\lambda)=\frac{2(n-1)\bar{p}-\frac{4}{n}\sum_{i<j}p_ip_j+2n(n-1)\gamma\bar{p}(1-\bar{p})}{\lambda}+\sum_k\frac{a_{k,c_i=1,\{c_j=0:j\neq i\}}}{\lambda-\lambda_k},$$

where $a_{k,c_i=1,\{c_j=0:j\neq i\}}$ depend on p_i, n, γ, and λ_k are eigenvalues of the generator (3.40). Applying the inverse transform to (3.59), we have

$$\mu^{(1)}_{c_i=1,\{c_j=0:j\neq i\}}(t)=2(n-1)\bar{p}-\frac{4}{n}\sum_{i<j}p_ip_j+2n(n-1)\gamma\bar{p}(1-\bar{p})+\tilde{\mu}^{(1)}_{c_i=1,\{c_j=0:j\neq i\}}(t), \tag{3.60}$$

where $\tilde{\mu}^{(1)}_{c_i=1,\{c_j=0:j\neq i\}}(t)$ represent terms decay as $t\to\infty$. Thus, we have

$$u^{(1)}(p_1,p_2,...,p_n)=2(n-1)\bar{p}-\frac{4}{n}\sum_{i<j}p_ip_j+2n(n-1)\gamma\bar{p}(1-\bar{p}). \tag{3.61}$$

3.6. Summary

Fixation of a single mutant under directional selection, where the mutant spreads in a multigene family by a gene conversion, is investigated. It is found that under genic selection the fixation probability is independent of rates of gene conversion and recombination, and initial linkage disequilibrium. By simulations Mano and Innan (2008) demonstrated that the formula could be extended to n loci. Interestingly, the formula (3.14) is given by the formula for the single locus fixation probability with replacing σ by $n\sigma$ and p by $\bar{p}$. The result can be interpreted as follows; gene conversion is a "migration" between loci and the effective size of the total population is enlarged to nN. In fact, the formula (3.12) already

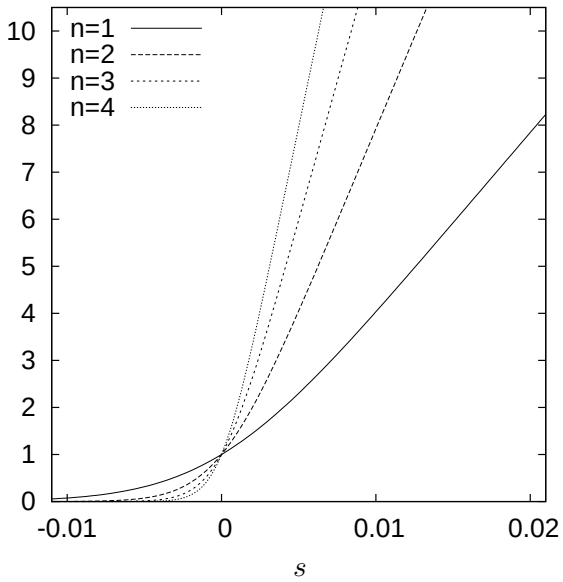

FIGURE 3.1. The evolutionary rate of duplicated genes under concerted evolution by gene conversion with directional selection.

appeared in a problem with population structure and migration (Maruyama, 1972). When dominance exists, the simple correspondence between the n-locus fixation probability and the single locus fixation probability no longer holds.

The formula (3.14) has interesting implications on the evolutionary rate of duplicated genes under selective pressure. Assume mutation appears at rate v per generation per locus. When a new mutant appears, we have $\bar{p} = 1/(2nN)$. Thus, the substitution rate is

$$2nNv \times \frac{1 - e^{-2s}}{1 - e^{-n\sigma}}. \tag{3.62}$$

The substitution rate (divided by v) with $N = 100$ is shown in Figure 3.1. It is shown that the formula implicates that selection works more efficiently in a larger gene family: the substitution rate of an advantageous mutant is higher for a larger family, while that of deleterious mutant is lower. Thus, it seems that having more copies in a family could enhance the action of directional selection. The result predicts that the evolutionary rate at nonsynonymous sites is different between a single-copy gene and multigene family, while this does not hold for the rate at synonymous sites. This prediction may well explain a recent report, which demonstrated that the substitution rate at nonsynonymous sites is accelerated in concerted evolving gene cluster in *Caenorhabditis elegens* and its relatives (Thomas, 2006).

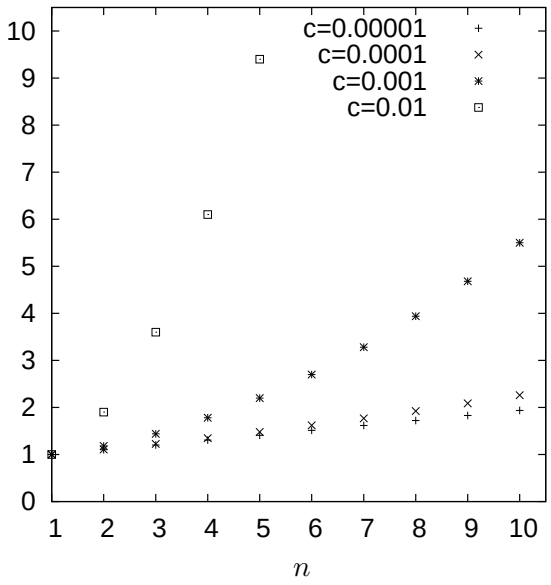

FIGURE 3.2. The evolutionary rate of duplicated genes under concerted evolution by biased gene conversion.

Then, fixation of a single mutant under weak bias in conversion rate, where the mutant spreads in a multigene family by a gene conversion, is investigated. The acceleration of the fixation probability by biased conversion is given by (3.49). The formula (3.49) also has interesting implications on the evolutionary rate of duplicated genes under biased gene conversion. Assume mutation appears at rate v per generation per locus. Without loss of generality, we may set $p_1 = 1/(2N), p_2 = p_3 = ... = p_n = 0$, and $\bar{p} = 1/(2nN)$. The fixation probability of the mutant is

$$u\left(\frac{1}{2N}, 0, 0, ..., 0\right) = \frac{1}{2nN} + \epsilon\left\{\frac{1}{N}\left(1 - \frac{1}{n}\right) + \frac{n-1}{N}\left(1 - \frac{1}{2nN}\right)\gamma\right\} + O(\epsilon^2). \tag{3.63}$$

Up to the first order in ϵ, the substitution rate is

$$2nNv \times u\left(\frac{1}{2N}, 0, 0, ..., 0\right) = v + 2(n-1)(1+n\gamma)v\epsilon + O(N^{-1}). \tag{3.64}$$

The substitution rate (divided by v) with $N = 100$ and $\epsilon = 0.05$ is shown in Figure 3.2. It is shown that when γ is small, the acceleration of the evolutionary rate proportional to size of the gene family. When γ is large, the acceleration of the evolutionary rate is proportional to square of size of the gene family. It seems that having more copies in a family could accelerate the substitution rate. Interestingly, the prediction is in accordance with the recent finding on GC content evolution in mammalian histone gene families. Galtier (2003) found highly significant correlation between the GC content at the third

codon positions and the copy number. The observation supports our theoretical prediction, as long as gene conversion is a biased process that tends to increase GC content.

CHAPTER 4

Ancestral Selection Graphs

4.1. Introduction

In population genetics, the ancestral genealogy of a sample of genes plays an important role in a probabilistic description of the sample. Consider a discrete-time Wright-Fisher model of a population consisting of $2N$ neutral genes. A diffusion time scaling is to measure time in units of $2N$ generations and let $2N \to \infty$. The Wright-Fisher model converges to a diffusion process. The process of sample of n genes' ancestors backward in time is described by the coalescent process (Kingman (1982b)). The convergence is robust under a number of different models (e.g. Moran model). Let $a_n(t)$ be the number of ancestors at time t backward of a sample of n neutral genes. Then the size process $\{a_n(t); t \geq 0\}$ is a death process with death rate $i(i-1)/2$ when the size is i. The size process will be referred to as the ancestral process. The distribution of $a_n(t)$ is known (see Griffiths (1979), Tavaré (1984)) and

$$\mathbb{P}[a_n(t) = i] = \sum_{k=i}^{n} \frac{(-1)^{k-i}(2k-1)(i)_{k-1}[n]_k}{i!(k-i)!(n)_k} \rho_k^0(t), \qquad i = 1, 2, ..., n, \tag{4.1}$$

where $\rho_k^0(t) := \exp\{-k(k-1)t/2\}$. The total variation norm between $a_n(t)$ and $a_n(\infty) = \delta_1$ has a simple form

$$\|a_n(t), \delta_1\|_{var} = 1 - \mathbb{P}[a_n(t) = 1]. \tag{4.2}$$

There is a first time $W_{n,1}^0$ such that $a_n(W_{n,1}^0) = 1$. The density of $W_{n,1}^0$ follows

$$\mathbb{P}[W_{n,1}^0 \leq t] = \mathbb{P}[a_n(t) = 1] = \sum_{k=1}^{n} \frac{(-1)^{k-1}(2k-1)[n]_k}{(n)_k} \rho_k^0(t). \tag{4.3}$$

A bound for $\mathbb{P}[a_n(t) = 1]$ is known (see Kingman (1982b), Tavaré (1984)) and

$$\rho_2^0(t) \leq 1 - \mathbb{P}[a_n(t) = 1] \leq 3\frac{n-1}{n+1}\rho_2^0(t), \qquad n = 2, 3, ... \tag{4.4}$$

The ancestral selection graph introduced by Krone and Neuhauser (1997) is an analogue of the coalescent genealogy. Assume that a pair of allelic types A_1 and A_2 are segregating in a population, and the selective advantage of a type A_1 gene over a type A_2 gene is s (> 0). Let $N \to \infty$ while $c = Ns$ is held constant. The elements are referred to

as particles. Let $b_n(t)$ be the number of edges, or ancestral particles, in a cross section of an ancestral selection graph at time t backward of a sample of n genes. In the ancestral selection graph, coalescing occurs at rate $\alpha_i = i(i-1)/2$ and branching occurs at rate $\beta_i = 2ci$ when the size is i. Then the ancestral process $\{b_n(t); t \geq 0\}$ is a birth and death process with rates β_i and α_i. A particle is called real if it is a part of the real genealogy of the sample, otherwise the particle is called virtual. If two particles reach a coalescing point, the resulting particle is real if and only if at least one of the two particles is real, otherwise the resulting particle is virtual. If a real particle reaches a branching point, it splits into a real particle and into a virtual particle. If a virtual particle reaches a branching point, it splits into two virtual particles. If a type A_2 particle reaches a branching point, it splits into two type A_2 particles. If a type A_1 particle reaches a branching point, it splits into two particles, where at least one of the two particles is type A_1. There is a first time $W^c_{n,1}$ such that $b_n(W^c_{n,1}) = 1$ because quadratic death and linear birth rates. Krone and Neuhauser (1997) consider stopping the process at this time, since the genetic composition of the sample is determined by then. They called the ancestral particle at the time the ultimate ancestor. In the case of no mutation, the real genealogy of a sample is the same as in the neutral process (Theorem 3.12 in Krone and Neuhauser (1997)). Thus the ancestral process of the real particles can be described by the neutral process $\{a_n(t); t \geq 0\}$, however, few properties of the ancestral process $\{b_n(t); t \geq 0\}$ are known. In this article, properties of the ancestral process $\{b_n(t); t \geq 0\}$ which is not stopped upon reaching the ultimate ancestor are studied. Fearnhead (2002) has studied a process which is not stopped upon reaching the ultimate ancestor. In particular, he identifies the stationaru distribution of this process and uses the distribution to characterize the substitution process to the common ancestor.

Kimura (1955c) studied the density of the allele frequency by the diffusion process to which the Wright-Fisher model with selection converges. Let $x_p(t)$ be the frequency of the allele A_1 at time t forward of a population in which the initial frequency of the allele A_1 is p. Then the Kolmogorov forward equation for the diffusion process $\{x_p(t); t \geq 0\}$ on $(0, 1)$ is

$$\frac{\partial \phi}{\partial t} = \frac{1}{2}\frac{\partial^2}{\partial x^2}\{x(1-x)\phi\} - 2c\frac{\partial}{\partial x}\{x(1-x)\phi\}, \tag{4.5}$$

with the initial condition $\phi(p, x; 0) = \delta(x - p)$. The solution is

$$\phi(p, x; t) = 2(1-r^2)e^{c(r-1)}e^{2cx}\sum_{k=0}^{\infty}\frac{V^{(1)}_{1k}(c, r)V^{(1)}_{1k}(c, z)}{N_{1k}}\rho^c_{k+2}(t), \tag{4.6}$$

where $r = 1 - 2p,\ z = 1 - 2x,\ \rho^c_{k+2}(t) := \exp(-\lambda_k t),\ k = 0, 1, 2, ...$ and $-\lambda_k$ $(0 < \lambda_0 < \lambda_1 < \cdots)$ are eigenvalues of the generator. $V^{(1)}_{1k}(c, z)$ is the oblate spheroidal wave function (see Appendix):

$$V^{(1)}_{1k}(c, z) = {\sum_{l\geq 0}}' f^k_l(c) T^1_l(z), \tag{4.7}$$

where $T^1_l(z)$ is the Gegenbauer function (may also be denoted by $C^{\frac{3}{2}}_l(z)$) and the summation is over even values of l if k is even, odd values of l if k is odd. N_{1k} is the normalization constant of $V^{(1)}_{1k}(c, z)$. The probability mass at the exit boundaries are

$$f(p, 1; t) = 2(1 - r^2)e^{c(r+1)} \sum_{k=0}^{\infty} \frac{V^{(1)}_{1k}(c, r)V^{(1)}_{1k}(c, -1)}{2\lambda_k N_{1k}}(1 - \rho^c_{k+2}(t)), \tag{4.8}$$

and

$$f(p, 0; t) = 2(1 - r^2)e^{c(r-1)} \sum_{k=0}^{\infty} \frac{V^{(1)}_{1k}(c, r)V^{(1)}_{1k}(c, 1)}{2\lambda_k N_{1k}}(1 - \rho^c_{k+2}(t)). \tag{4.9}$$

In Section 4.2, the ancestral process $\{b_n(t); t \geq 0\}$ without absorbing states is studied. An explicit form of the probability distribution of $b_n(t)$ are obtained, by using a dual relationship between the ancestral process and the diffusion process in a context by Tavaré (1984). The ancestral process converges to the stationary measure. In Section 4.3, the convergence and bounds are discussed. In contrast to the neutral process, the final rates of convergence are given by the largest eigenvalue for all the states. Bounds for the probability that $b_n(t)$ is at the state 1 are obtained by an elementary martingale argument, which corresponds to the bounds (4.4) for the neutral process. In Section 4.4, the ancestral process with absorbing states are considered. It is shown that the first passage times of the ancestral process $\{b_n(t); t \geq 0\}$ at the states $1, 2, ..., n - 1$ are larger than that in the neutral process for all the states. By killing the modified process, in which the state 1 is the absorbing state, the formal form of the joint probability generating function of $b_n(t)$ and the number of branching events is obtained. By using the formula, the expectation of the total length of the edges in the ancestral selection graph is obtained. In Section 4.5, the ancestral process of the whole population $\{b_\infty(t); t \geq 0\}$ is studied. It is shown that the process of fixation of the allele in the diffusion model corresponds to convergence of the ancestral process to its stationary measure. The time to fixation of an allele conditional on fixation is studied in terms of the ancestral process. It is shown that the density of time to fixation of a single mutant gene conditional on fixation is given by the probability of the whole population being descended from a single real ancestral particle, regardless

of the allelic type. In the neutral process, the density of the waiting time to the ancestral process hits the state 1 and the density of the conditional fixation time are given by the probability that the ancestral process is at the state 1. The property does not hold in the process with selection.

4.2. Number of ancestral particles

In this section, we will obtain an equation that relates the moments of the Wright-Fisher diffusion with directional selection to a Markov process that specifies the number of particles (real and virtual) that are present in the ancestral selection graph. To derive this result, we will exploit the concept of duality from the theory of Markov processes (Ethier and Kurtz, 1986). If $X = \{X_t; t \geq 0\}$ and $Y = \{Y; t \geq 0\}$ are Markov processes with values in sets Z_X and Z_Y, respectively, then X and Y are said to be dual with respect to a function $f(x, y)$ if the identity

$$\mathbb{E}_x[f(X_t, y)] = \mathbb{E}_y[f(x, Y_t)] \tag{4.10}$$

holds for every $x \in Z_X$ and $y \in Z_Y$. Duality is a useful concept because it allows us to use our knowledge of one process to learn about the other. Although there is no general procedure for identifying dual processes, duality can sometimes be deduced using simple generator calculations. Specifically, if G_x is the infinitesimal generator of the process X and G_y is the infinitesimal generator of the process Y, then the duality relationship shown in (4.10) will be satisfied if the identity

$$G_x f(x, y) = G_y f(x, y) \tag{4.11}$$

holds for all x and y. Here we think $G_x f(x, y)$ as acting on the x-variable of the function $f(x, y)$ for each fixed value of y.

To apply these results to the Wright-Fisher diffusion with selection, it will be necessary to consider the frequency $y_q(t)$ of the less fit allele, which is itself governed by a Wright-Fisher diffusion with generator:

$$G_y f(y) = \frac{1}{2} y(1-y) f''(y) - 2cvy(1-y) f'(y). \tag{4.12}$$

Notice that the selection coefficient is negative in this case ($c \geq 0$). If we define the function $f(y, n) = y^n$, then a simple calculation shows that

$$\begin{aligned} G_y f(y, n) &= \binom{n}{2} [f(y, n-1) - f(y, n)] + 2cn[f(y, n+1) - f(y, n)] \\ &= G_n f(y, n), \end{aligned} \tag{4.13}$$

where G_n is the operator defined by

$$G_n f(n) = \binom{n}{2}[f(n-1) - f(n)] + 2cn[f(n+1) - f(n)]. \tag{4.14}$$

In other words, G_n is the infinitesimal generator of the birth-death process $\{b_n(t); t \geq 0\}$ which keeps track of the number of ancestral particles in the ancestral selection graph. Because $f(y, n)$ is bounded, we can use a result of Ethier and Kurtz (1986) to deduce that the Wright-Fisher diffusion $\{y_q(t); t \geq 0\}$ and the birth-death process $b_n(t)$ are dual with respect to the function $f(y, n)$:

THEOREM 4.2.1.

$$\mathbb{E}[q^{b_n(t)}] = \mathbb{E}[(y_q(t))^n], \qquad n = 1, 2, ... \tag{4.15}$$

Because the right-hand side of this identity involves moment of the process $\{y_q(t); t \geq 0\}$, the process $\{b_n(t); t \geq 0\}$ is said to be a moment dual for $\{y_q(t); t \geq 0\}$. The existence of moment duals of Wright-Fisher diffusions with polynomial coefficients was first shown by Shiga (1981), and the explicit description of duality between the birth-death process and the Wright-Fisher diffusion with directional selection is discussed in Athreya and Swart (2005). By the Itô formula, we have a system of differential equations for the moments of $y_q(t)$

$$\frac{d\xi_n}{dt} = -(\alpha_n + \beta_n)\xi_n + \alpha_n \xi_{n-1} + \beta_n \xi_{n+1}, \qquad n = 1, 2, ... \tag{4.16}$$

where $\xi_n = \mathbb{E}[(y_q(t))^n]$. The Kolmogorov backward equation for the ancestral process $\{b_n(t); t \geq 0\}$ without absorbing states is also given by (4.16), where $\xi_n = \mathbb{P}[b_n(t) = i]$. Thus, (4.16) with $\xi_n = \mathbb{E}[q^{b_n(t)}]$ holds. The isomorphism of these equations is a consequence of (4.15). A proof of (4.15) by using graph theoretical arguments is also possible.

PROOF. Partition an ancestral selection graph $\mathcal{G}$ into disconnected subgraphs $\mathcal{G}_i, i = 1, 2, ...$ Let $\mathcal{E}_t$ be the edges, or the ancestral particles, of a cross section of $\mathcal{G}$ taken at time t backward. Then, $b_n(t) = |\mathcal{E}_t|$. Each $\mathcal{E}_0 \cap \mathcal{G}_i$ consists only of type A_2 particles if and only if $\mathcal{E}_t \cap \mathcal{G}_i$ consists only of type A_2 particles, since at least one type A_1 particle survives from time t to 0 if $\mathcal{E}_t \cap \mathcal{G}_i$ contain type A_1 particles. Here the ancestral selection graph is viewed forward in time. If a type A_1 particle reaches a coalescing point, the number of type A_1 particles increase by 1. If a type A_1 particle reaches a branching point and meets other particle, the resulting particle is always type A_1. Thus, a sample consists only of

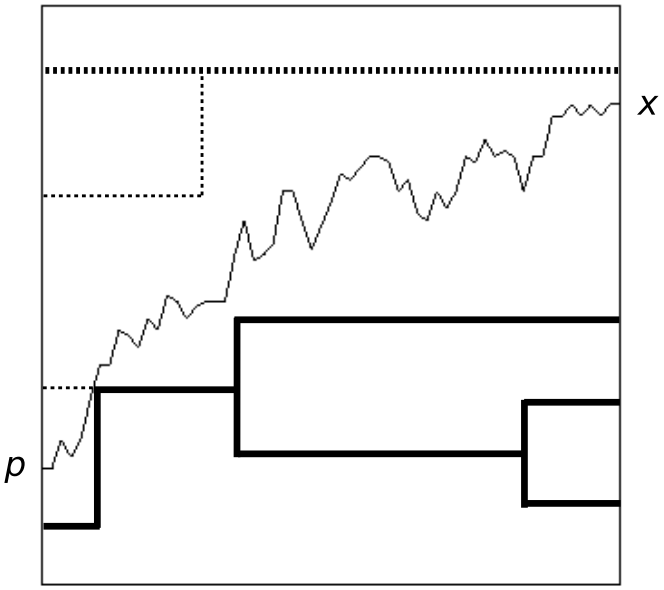

FIGURE 4.1. A realization of the ancestral selection graph embedded in a diagram of a sample path of $x_p(t)$.

type A_2 particles if and only if the ancestral particles at time t consists only of type A_2 particles. □

If a sample contains type A_1 particles, then for $n = 0, 1, ...; m = 1, 2, ...,$

$$\mathbb{E}[(x_p(t))^m (y_q(t))^n] = \mathbb{E}[(1 - y_q(t))^m (y_q(t))^n] = \sum_{i=0}^{m} \frac{(-1)^i m!}{i!(m-i)!} \mathbb{E}[q^{b_{i+n}(t)}], \tag{4.17}$$

with the convention $b_0(t) = 0$. In particular,

$$\mathbb{E}[x_p(t)] = \mathbb{E}[1 - y_q(t)] = \mathbb{E}[1 - q^{b_1(t)}] = \sum_{k=1}^{\infty} \mathbb{P}[b_1(t) = k] \sum_{i=1}^{k} \frac{k! p^i q^{k-i}}{i!(k-i)!}. \tag{4.18}$$

The expression follows immediately from a distribution of particles in $\mathcal{E}_t$. Since the ancestral selection graph is irreducible, an type A_1 particle is sampled if and only if $\mathcal{E}_t$ contains at least one type A_1 particle. The first summation is over $|\mathcal{E}_t|$ and the second summation is over the number of type A_1 particles.

A realization of the ancestral selection graph consists of two disconnected subgraphs embedded in a diagram of a sample path of $x_p(t)$ can be seen in Figure 1. The abscissa is time interval $(0, t)$ and ordinate is $x_p(t)$. Thick lines represent the real genealogy. Lines used by type A_2 particles are dotted. The graph contributes to $\mathbb{E}[(x_p(t))^3 y_q(t)]$ and $b_4(t) = 4$. If an ancestral selection graph consists of the upper subgraph only, it contributes to $\mathbb{E}[y_q(t)] = \mathbb{E}[q^{b_1(t)}]$ and $b_1(t) = 2$.

Using an integral transform by the Gegenbauer function (Erdélyi (1954)) for $l = 0, 1, ...; n = 1, 2, ...; i = 0, 1, ...,$

$$\int_{-1}^{1} T_l^1(z)(1+z)^n(1-z)^i dz = \frac{2^{n+i} i!(l+1)(l+2)}{(n+1)_{i+1}} {}_3F_2(-l, l+3, i+1; 2, i+n+2; 1), \tag{4.19}$$

where ${}_3F_2(-l, l+3, i+1; 2, i+n+2; 1)$ is the generalized hypergeometric function, and with an identity (4.77), it is possible to obtain explicit form of the probability generating function of $b_n(t)$, and we have

$$\begin{aligned}
\mathbb{E}[q^{b_n(t)}] &= \mathbb{E}[y_q(t)^n] = \int_0^1 (1-x)^n \phi(x, p; t) dx + f(p, 0; t) \\
&= \frac{e^{4cq} - 1}{e^{4c} - 1} + 2(1-r^2) e^{c(r-1)} \sum_{k=0}^{\infty} \frac{V_{1k}^{(1)}(c, r)}{N_{1k}} \left\{ F_{kn}(c) - \frac{V_{1k}^{(1)}(c, 1)}{2\lambda_k} \right\} \rho_{k+2}^c(t) \\
&= \sum_{i=1}^{\infty} \mathbb{P}[b_n(t) = i] q^i, \qquad (4.20)
\end{aligned}$$

where

$$F_{kn}(c) := {\sum_{l \geq 0}}' f_l^k(c) \sum_{i=0}^{\infty} \frac{(2c)^i}{(n+1)_{i+1}} \sum_{j=0}^{l} \frac{(-l)_j (l+1)_{j+2} (i+1)_j}{2 \cdot j! (j+1)! (i+n+2)_j}.$$

If $n = \infty$, $f(p, 0; t)$ gives the probability generating function. Using a power series expansion in q of the Gegenbauer function

$$T_l^1(r) = (-1)^l \sum_{i=0}^{l} \frac{(-l)_i (l+1)_{i+2}}{2 \cdot i! (i+1)!} q^i, \qquad l = 0, 1, .. \tag{4.21}$$

we obtain explicit form of the probability distribution of $b_n(t)$:

$$\mathbb{P}[b_n(t) = 1] = \pi_1 + 8e^{-2c} \sum_{k=0}^{\infty} \frac{V_{1k}^{(1)}(c, -1)}{N_{1k}} \left\{ F_{kn}(c) - \frac{V_{1k}^{(1)}(c, 1)}{2\lambda_k} \right\} \rho_{k+2}^c(t), \tag{4.22}$$

and

$$\mathbb{P}[b_n(t) = i] = \pi_i + 8e^{-2c} \sum_{k=0}^{\infty} \frac{G_{ki}(c)}{N_{1k}} \left\{ F_{kn}(c) - \frac{V_{1k}^{(1)}(c, 1)}{2\lambda_k} \right\} \rho_{k+2}^c(t), \qquad i = 2, 3, ..., \tag{4.23}$$

where

$$\begin{aligned}
G_{ki}(c) &:= {\sum_{l \geq i-1}}' f_l^k(c) (-1)^l \frac{(-l)_{i-1} (l+1)_{i+1}}{2(i-1)! i!} \\
&\quad + \sum_{j=1}^{i-1} \frac{(2c)^{j-1}(2c-j)}{j \cdot (j-1)!} {\sum_{l \geq i-j-1}}' f_l^k(c) (-1)^l \frac{(-l)_{i-j-1} (l+1)_{i-j+1}}{2(i-j-1)!(i-j)!},
\end{aligned}$$

and π_i are given in (4.33). Note that there are finite probabilities at the states $n+1, n+2, ...$ The expected number of ancestral particles is

$$\begin{aligned}\mathbb{E}[b_n(t)] &= \pi_1 e^{4c} + 8e^{-2c}\sum_{k=0}^{\infty}\frac{V_{1k}^{(1)}(c,-1)}{N_{1k}}\left\{F_{kn}(c) - \frac{V_{1k}^{(1)}(c,1)}{2\lambda_k}\right\}\rho_{k+2}^{c}(t) \\ &\quad +8e^{-2c}\sum_{i=2}^{\infty} i\sum_{k=0}^{\infty}\frac{G_{ki}(c)}{N_{1k}}\left\{F_{kn}(c) - \frac{V_{1k}^{(1)}(c,1)}{2\lambda_k}\right\}\rho_{k+2}^{c}(t), \qquad (4.24)\end{aligned}$$

and the falling factorial moments are

$$\mathbb{E}[[b_n(t)]_i] = i!\pi_i e^{4c} + 8e^{-2c}\sum_{j=i}^{\infty}[j]_i\sum_{k=0}^{\infty}\frac{G_{kj}(c)}{N_{1k}}\left\{F_{kn}(c) - \frac{V_{1k}^{(1)}(c,1)}{2\lambda_k}\right\}\rho_{k+2}^{c}(t), \qquad i = 2,3,... \qquad (4.25)$$

For small c, the probability distribution is approximately

$$\mathbb{P}[b_n(t)=1] = \mathbb{P}[a_n(t)=1] - 2c + 2c\sum_{k=2}^{n+1}(-1)^k(2k-1)\left\{\frac{[n]_k}{(n)_k} + \frac{k(k-1)[n]_{k-1}}{(n)_{k+1}}\right\}\rho_k^0(t) + O(c^2), \qquad (4.26)$$

and for $i = 2, 3, ...,$

$$\begin{aligned}\mathbb{P}[b_n(t)=i] &= \mathbb{P}[a_n(t)=i] + 2c\delta_{i,2} - 2c\sum_{k=i}^{n+1}\frac{(-1)^{k-i}(2k-1)(i)_{k-1}}{i!(k-i)!}\left\{\frac{k(k-1)[n]_{k-1}}{(n)_{k+1}}\right. \\ &\quad \left. +\frac{(k^2-k+2i-2)[n]_k}{(k-i+1)(k+i-2)(n)_k}\right\}\rho_k^0(t) + 2c\frac{(i-1)_{i-1}[n]_{i-1}}{(i-1)!(n)_{i-1}}\rho_{i-1}^0(t) + O(c^2), \qquad (4.27)\end{aligned}$$

with a convention $[n]_{n+1} = 0$.

It is possible to obtain the solution of (4.16) as a perturbation series in $2c$, where the series is represented by eigenvalues of the neutral process. Let

$$\xi_n = \xi_n^{(0)} + 2c\xi_n^{(1)} + (2c)^2\xi_n^{(2)} + \cdots, \qquad n = 1,2,... \qquad (4.28)$$

Denote the infinitesimal generator of the neutral process $\{a_n(t); t \geq 0\}$ by $Q_0 = (q_{0,ij})$, where $q_{0,i+1,i} = \alpha_{i+1}, q_{0,ii} = -\alpha_i$ for $i = 1,2,...$ and other elements are zero. Let the Laplace transform of $\xi_n^{(i)}(t)$ be $\nu_n^{(i)}(\lambda)$. It is straightforward to show that

$$\nu^{(i)} = \{(Q_0 - \lambda E)^{-1}C\}^i(\lambda E - Q_0)^{-1}\xi^{(0)}(0), \qquad i = 1,2,... \qquad (4.29)$$

where $C = (c_{ij})$ is given by $c_{ii} = i, c_{i,i+1} = -i$ for $i = 1,2,...$ and other elements are zero. Note that the inverse Laplace transform of the element in the n-th row and i-th column of the matrix $\{(Q_0 - \lambda E)^{-1}C\}^j(\lambda E - Q_0)^{-1}$ gives the j-th order coefficients in $2c$ of $\mathbb{P}[b_n(t) = i]$.

Let $r_n(t)$ be the number of branching events in the time interval $(0,t)$ in the ancestral selection graph of a sample of n genes, where $r_n(0)=0$. The joint probability generating functions of $b_n(t)$ and $r_n(t)$ satisfy a system of differential equation

$$\frac{d\xi_n}{dt} = -(\alpha_n + v\beta_n)\xi_n + \alpha_n\xi_{n-1} + v\beta_n\xi_{n+1} - (1-v)\beta_n\xi_n, \qquad n = 1, 2, .. \tag{4.30}$$

with the initial condition $\xi_n(0) = q^n$, where $\xi_n = \mathbb{E}[q^{b_n(t)}u^{r_n(t)}]$. The formal solution is given by killing of the modified process $\{\tilde{b}_n(t); t \geq 0\}$ in which the selection coefficient is vc, and we have

$$\mathbb{E}[q^{b_n(t)}v^{r_n(t)}] = \mathbb{E}\left[q^{\tilde{b}_n(t)} \exp\left\{-2c(1-v)\int_0^t \tilde{b}_n(u)du\right\}\right]. \tag{4.31}$$

It is straightforward to show that

$$\mathbb{E}[q^{b_n(t)}, r_n(t) = 0] = \mathbb{E}\left[q^{a_n(t)} \exp\left\{-2c\int_0^t a_n(u)du\right\}\right], \tag{4.32}$$

which shows the Poisson nature of the branching in the ancestral process $\{b_n(t); t \geq 0\}$. The integral is the total length of the edges in the neutral genealogy without branching in $(0,t)$. In particular, $\mathbb{P}[b_1(t) = 1, r(t) = 0] = e^{-2ct}$.

4.3. Convergence and bounds

Standard results on birth and death processes (see, e.g., Karlin and Taylor (1975)) gives the stationary measure of the ancestral process $\{b_n(t); t \geq 0\}$. It is straightforward to obtain the stationary measure

$$\pi_i := \mathbb{P}[b_n(\infty) = i] = \frac{(4c)^i}{i!(e^{4c}-1)}, \quad i = 1, 2, ..., \tag{4.33}$$

which is the zero-truncated Poisson distribution. Since the ancestral process of the real particles is the neutral process $\{a_n(t); t \geq 0\}$ and the number of real particles becomes 1 in finite time, π_1 is the probability that there are no virtual particles.

It is clear from (4.22) and (4.23) that for $i = 1, 2...$,

$$\pi_i - \mathbb{P}[b_n(t) = i] = O(\rho_2^c(t)), \qquad t \to \infty. \tag{4.34}$$

For small c, the final rates of convergence are approximately

$$\lim_{t\to\infty} (\rho_2^c(t))^{-1} \{\pi_1 - \mathbb{P}[b_n(t) = 1]\} = 3e^{-2c}\left\{\frac{n-1}{n+1} - \frac{4c}{(n+1)(n+2)}\right\} - 2ce^{-2c}\delta_{n,1} + O(c^2), \tag{4.35}$$

$$\lim_{t\to\infty} (\rho_2^c(t))^{-1} \{\pi_2 - \mathbb{P}[b_n(t) = 2]\} = -3e^{-2c}\left\{\frac{n-1}{n+1} - \frac{2n(n-1)c}{n+2}\right\} - 2ce^{-2c}\delta_{n,1} + O(c^2), \tag{4.36}$$

and

$$\lim_{t\to\infty}(\rho_2^c(t))^{-1}\{\pi_i - \mathbb{P}[b_n(t)=i]\} = -3e^{-2c}\frac{n-1}{n+1}\frac{(2c)^{i-2}}{(i-2)!} + O(c^{i-1}), \qquad i=3,4,... \tag{4.37}$$

In contrast to the neutral process, the final rates of convergence are given by the largest eigenvalue for all the states. In the neutral process, we have

$$\lim_{t\to\infty}(\rho_i^0(t))^{-1}\mathbb{P}[a_n(t)=i] = \frac{(i)_i[n]_i}{i!(n)_i}, \qquad i=1,2,...,n. \tag{4.38}$$

The total variation norm has no simple form as in (4.2).

A simple argument gives a bound for $\mathbb{P}[b_n(t)=1]$. The event that the number of ancestral particles is 1 is a subset of the event that the number of real particle is 1, and we have

$$\mathbb{P}[b_n(t)=1] \le \mathbb{P}[a_n(t)=1], \qquad n=1,2,... \tag{4.39}$$

An elementary argument on a martingale gives bounds for $\mathbb{P}[b_n(t)=1]$ directly. Let $\eta(n;c)$ satisfy a recursion

$$(\lambda_0-\alpha_n-\beta_n)\eta(n;c) + \alpha_n\eta(n-1;c) + \beta_n\eta(n+1;c) = 0, \qquad n=1,2,... \tag{4.40}$$

with the boundary condition $\eta(1;c) = -2c$. Since η is an eigenvector of the transition probability matrix of the ancestral process $\{b_n(t); t\ge 0\}$, $\eta(b_n(t);c)(\rho_2^c(t))^{-1}$ is a martingale to the ancestral process (see, e.g., Karlin and Taylor (1975)). Then,

$$\mathbb{E}[\eta(b_n(t);c)] = \eta(n;c)\rho_2^c(t). \tag{4.41}$$

Although the explicit form of $\eta(n;c)$ is not available, it is possible to obtain an asymptotic form. Because of

$$\frac{\eta(n;c)}{\eta(n-1;c)} \to 1 + \frac{2\lambda_0}{n(n-1)} + O(n^{-3}), \qquad n\to\infty, \tag{4.42}$$

we deduce the asymptotic form $\eta(n;c) \approx \eta(\infty;c)(1-2\lambda_0/n)$, where $\eta(\infty;c)$ is a function of c. Then, the derivative in c has an asymptotic form

$$\log\eta'(n;c) \approx 4c\frac{\{\eta'(\infty;c)+\eta(\infty;c)\}^2}{\eta'(\infty;c)^2}\log n, \qquad n\to\infty. \tag{4.43}$$

If $\eta(n;c)$ is finite, then $\eta'(\infty;c)+\eta(\infty;c)=0$. For the neutral process $\{a_n(t); t\ge 0\}$, it is known that $\eta(\infty;0)=3$ (see, Kingman (1982a)). Thus, $\eta(\infty;c) = 3e^{-c}$.

LEMMA 4.3.1. *$\eta(n;c)$ is monotonically increasing in n.*

PROOF. By taking $t = \infty$ in (4.41) it follows that

$$\sum_{i=1}^{\infty} \eta(i;c)\pi_i = 0. \tag{4.44}$$

Denote the infinitesimal generator of the ancestral process $\{b_n(t); t \geq 0\}$ by $Q_c = (q_{c,ij})$, where $q_{c,i+1,i} = \alpha_{i+1}, q_{c,ii} = -(\alpha_i + \beta_i), q_{c,i,i+1} = \beta_i$ for $i = 1, 2, ...$ and other elements are zero. η is an eigenvector of an oscillatory matrix $E+Q_c(2N)^{-1}$ which belongs to the second largest eigenvalue $1 - \lambda_0(2N)^{-1}$. An eigenvector of an oscillatory matrix which belongs to the second largest eigenvalue has exactly one variation of sign in the coordinates (see, Gantmacher (1959), pp. 105). Assume $\eta(i;c) > 0,\ i \geq L$ and $\eta(i;c) \leq 0,\ 1 \leq i \leq L-1$. Suppose $l \geq L-1$. By an induction we deduce from (4.40) that

$$\begin{aligned} \eta(l+1;c) - \eta(l;c) &= \frac{\alpha_l}{\beta_l}\{\eta(l;c) - \eta(l-1;c)\} - \lambda_0\eta(l;c) \\ &= \frac{(l-1)!}{(4c)^{l-1}}\left\{\eta(2;c) - \eta(1;c) - \frac{\lambda_0}{\pi_2}\sum_{i=2}^{l}\eta(i;c)\pi_i\right\} \\ &= \frac{\lambda_0}{8c^2\pi_{l-1}}\sum_{i=l+1}^{\infty}\eta(i;c)\pi_i > 0. \end{aligned} \tag{4.45}$$

Next, suppose $2 \leq l \leq L-2$. We have

$$\eta(l+1;c) - \eta(l;c) = \frac{(l-1)!}{(4c)^{l-1}}\left\{\eta(2;c) - \eta(1;c) - \frac{\lambda_0}{\pi_2}\sum_{i=2}^{l}\eta(i;c)\pi_i\right\} > 0. \tag{4.46}$$

Finally, $\eta(2;c) - \eta(1;c) = \lambda_0 > 0$. □

From Lemma 4.3.1 it follows that

$$\begin{aligned} \mathbb{P}[b_n(t) = 1]\eta(1;c) + \mathbb{P}[b_n(t) > 1]\eta(2;c) &\leq \mathbb{E}[\eta(b_n(t);c)] \\ &\leq \mathbb{P}[b_n(t) = 1]\eta(1;c) + \mathbb{P}[b_n(t) > 1]\eta(\infty;c), \end{aligned} \tag{4.47}$$

here we note that there are finite probabilities at the states larger than n. Then, from (4.41) we have the following bounds:

THEOREM 4.3.2. *If $\eta(n;c)$ satisfies the recursion (4.40), then*

$$\frac{\eta(n;c)\rho_2^c(t) + 2c}{3e^{-c} + 2c} \leq 1 - \mathbb{P}[b_n(t) = 1] \leq \frac{\eta(n;c)\rho_2^c(t) + 2c}{\lambda_0}, \qquad n = 1, 2, ... \tag{4.48}$$

REMARK 4.3.3. *For small c, $\eta(n;c)$ can be expanded into a power series in c.*

$$\eta(n;c) = 3\frac{n-1}{n+1}\left\{1 - c\frac{n^2+n+2}{(n-1)(n+2)}\right\} + O(c^2), \qquad n = 3, 4, ... \tag{4.49}$$

As $c \to \infty$, for $n = 1, 2, ...,$

$$\mathbb{P}[b_n(t) = 1] \to 0, \qquad t > 0. \tag{4.50}$$

PROOF. For small c, it is straightforward to obtain the power series expansion. For large c, the bounds are approximately

$$\inf_{n\geq 1}\left\{\frac{\eta(n;c)\rho_2^c(t)+2c}{3e^{-c}+2c}\right\} = 1 - \frac{3e^{-c}}{2c} - e^{-\lambda_0 t} + O(c^{-2}e^{-2c}), \tag{4.51}$$

and

$$\sup_{n\geq 1}\left\{\frac{\eta(n;c)\rho_2^c(t)+2c}{\lambda_0}\right\} = 1 + \frac{1}{2c} + \frac{3e^{-(c+\lambda_0 t)}}{2c} + O(c^{-2}), \tag{4.52}$$

where $\lambda_0 = 2c - 1 + O(c^{-2})$ (see Appendix). Then,

$$\lim_{c\to\infty}\inf_{n\geq 1}\left\{\frac{\eta(n;c)\rho_2^c(t)+2c}{3e^{-c}+2c}\right\} = \lim_{c\to\infty}\sup_{n\geq 1}\left\{\frac{\eta(n;c)\rho_2^c(t)+2c}{\lambda_0}\right\} = 1, \qquad t > 0. \tag{4.53}$$

□

COROLLARY 4.3.4. *For the whole population ($n = \infty$), the bounds reduce to*

$$\frac{3e^{-c}\rho_2^c(t)+2c}{3e^{-c}+2c} \leq 1 - \mathbb{P}[b_\infty(t) = 1] \leq \frac{3e^{-c}\rho_2^c(t)+2c}{\lambda_0}, \qquad n = 1, 2, ... \tag{4.54}$$

4.4. First passage times

Let

$$W_{n,i}^c := \inf\{t \geq 0; b_n(t) = i\}, \qquad i = 1, 2, ... \tag{4.55}$$

and $\{b_n^1(t); W_{n,1}^c \geq t \geq 0\}$ be a modified process, where there is an absorbing state at 1, or the ultimate ancestor. The modified process is the same as that introduced for ancestral recombination graph, where $4c$ is replaced by the recombination parameter ρ (Griffiths (1991)). Theorems 1, 2, 3 in Griffiths (1991) hold for the modified process. The modified process was studied by Krone and Neuhauser (1997). Here, modified processes $\{b_n^i(t); W_{n,i}^c \geq t \geq 0\}$, where there is an absorbing state at $i = 1, 2, ..., n-1$, are studied to discuss the first passage times of the ancestral process $\{b_n(t); t \geq 0\}$ at the states $1, 2, ..., n-1$.

It is possible to show that the expected first passage times of the ancestral process $\{b_n(t); t \geq 0\}$ at the states $1, 2, ..., n-1$ are larger than those in the neutral process $\{a_n(t); t \geq 0\}$. $\mathbb{E}[W_{n,1}^c]$ is given in Krone and Neuhauser (1997).

THEOREM 4.4.1. *Let*

$$W_{n,i}^0 := \inf\{t \geq 0; a_n(t) = i\}, \qquad i = 1, 2, ... \tag{4.56}$$

Then,

$$\mathbb{E}[W_{n,i}^c] = 2\sum_{k=i}^{n-1}\sum_{j=0}^{\infty}\frac{(4c)^j}{(k)_{j+2}} > \mathbb{E}[W_{n,i}^0], \qquad i = 1, 2, ..., n-1, \tag{4.57}$$

where $W_{n,1}^c$ is the time to the ultimate ancestor.

PROOF. The theorem follows from standard results on birth and death processes (see, e.g., Karlin and Taylor (1975)). The modified processes $\{b_n^i(t); W_{n,i}^c \geq t \geq 0\}$ hit the states $i = 1, 2, ..., n-1$ in finite time with probability one, since

$$\sum_{m=i}^{\infty}\prod_{k=i+1}^{m+1}\frac{\alpha_k}{\beta_k} = \frac{(4c)^{i-1}}{(i-1)!}\sum_{m=i}^{\infty}\frac{m!}{(4c)^m} = \infty, \qquad i = 1, 2, ..., n-1. \tag{4.58}$$

From the Kolmogorov backward equation for the modified process $\{b_n^i(t); W_{n,i}^c \geq t \geq 0\}$, which is (4.16) for $n = i+1, i+2, ...$ with $\xi_n = \mathbb{P}[b_n^i(t) = i]$ and the boundary condition $\xi_i = \delta(t)$, the expected first passage times satisfy a recursion for $i = 1, 2, ..., n-1$

$$(\alpha_n + \beta_n)\zeta(n) - \alpha_n\zeta(n-1) - \beta_n\zeta(n+1) = 1, \qquad n = i+1, i+2, ..., n-2 \tag{4.59}$$

with the boundary condition $\zeta(i) = 0$, where $\zeta(n) = \mathbb{E}[W_{n,i}^c]$. It is straightforward to solve the recursion and obtain

$$\mathbb{E}[W_{n,i}^c] = \sum_{m=i}^{\infty}\gamma_m + \sum_{j=i}^{n-2}\prod_{k=i+1}^{j+1}\frac{\alpha_k}{\beta_k}\sum_{l=j+1}^{\infty}\gamma_l = 2\sum_{k=i}^{n-1}\sum_{j=0}^{\infty}\frac{(4c)^j}{(k)_{j+2}}, \qquad i = 1, 2, ..., n-2, \tag{4.60}$$

and

$$\mathbb{E}[W_{n,n-1}^c] = \sum_{m=i}^{\infty}\gamma_m = 2\sum_{j=0}^{\infty}\frac{(4c)^j}{(k)_{j+2}}, \tag{4.61}$$

where

$$\gamma_i = \frac{1}{\alpha_{i+1}} = \frac{2}{i(i+1)}, \qquad \gamma_m = \frac{\beta_{i+1}\beta_{i+2}\cdots\beta_m}{\alpha_{i+1}\alpha_{i+2}\cdots\alpha_m\alpha_{m+1}} = \frac{2(4c)^{m-i}}{(i)_{m-i+2}}, \qquad m = i+1, i+2, ...$$

It is clear from (4.60) and (4.61) that

$$\mathbb{E}[W_{n,i}^c] > 2\sum_{k=i}^{n-1}\frac{1}{k(k+1)} = 2\left(\frac{1}{i} - \frac{1}{n}\right) = \mathbb{E}[W_{n,i}^0], \qquad i = 1, 2, ..., n-1. \tag{4.62}$$

□

As $c \to \infty$, for $i = 1, 2, ..., n-1$,

$$\mathbb{E}[W^c_{n,i}] = 2\sum_{k=i}^{n-1} \frac{1}{k(k+1)} \sum_{j=0}^{\infty} \frac{\left(\frac{4c}{k+2}\right)^j}{\prod_{l=0}^{j-1}\left(1+\frac{l}{k+2}\right)} > 2\sum_{k=i}^{n-1} \frac{e^{\frac{4c}{k+2}}}{k(k+1)} \to \infty. \tag{4.63}$$

COROLLARY 4.4.2. *For the whole population ($n = \infty$), the expected first passage times are*

$$\mathbb{E}[W^c_{\infty,i}] = 2\sum_{j=0}^{\infty} \frac{(4c)^j}{(j+1)(i)_{j+1}}, \qquad i = 1, 2, ... \tag{4.64}$$

PROOF. It follows immediately from an identity

$$\sum_{k=i}^{\infty} \frac{1}{(k)_{j+2}} = \frac{1}{(j+1)(i)_{j+1}}, \qquad j = 0, 1, ... \tag{4.65}$$

□

It is straightforward to obtain higher moments of the first passage times of the ancestral process $\{b_n(t); t \geq 0\}$ at the states $1, 2, ..., n-1$ in the same manner. The second moments $\mathbb{E}[(W^c_{n,i})^2]$ satisfy a recursion

$$(\alpha_n + \beta_n)\zeta(n) - \alpha_n\zeta(n-1) - \beta_n\zeta(n+1) = 2\mathbb{E}[W^c_{n,i}], \qquad n = i+1, i+2, ..., n-2 \tag{4.66}$$

with the boundary condition $\zeta(i) = 0$, where $\zeta(n) = \mathbb{E}[(W^c_{n,i})^2]$. However, there is no simple form for the density as in (4.3). The Laplace transform of the first passage times of the ancestral process satisfy a recursion for $i = 1, 2, ..., n-1$

$$(\lambda + \alpha_n + \beta_n)\zeta(n) - \alpha_n\zeta(n-1) - \beta_n\zeta(n+1) = 0, \qquad n = i+1, i+2, ..., n-2 \tag{4.67}$$

with the boundary condition $\zeta(i) = 1$, where $\zeta(n) = \mathbb{E}[e^{-\lambda W^c_{n,i}}]$.

The joint probability generating function of $b^1_n(t)$ and $r_n(t)$ satisfies a system of differential equation (4.30) with $\xi_n = \mathbb{E}[q^{b^1_n(t)} v^{r_n(t)}]$. By taking $t = \infty$, we have

$$0 = -(\alpha_n + \beta_n)\xi_n + \alpha_n\xi_{n-1} + v\beta_n\xi_{n+1}, \qquad n = 1, 2, .., \tag{4.68}$$

with the boundary condition $\xi_1 = 1$, where $\xi_n = \mathbb{E}[v^{r_n(\infty)}]$. The formal form of the probability generating function of $r(\infty)$ is

$$\mathbb{E}[v^{r_n(\infty)}] = \mathbb{E}\left[\exp\left\{-2c(1-v)\int_0^{W^{vc}_{n,1}} \tilde{b}^1_n(u)du\right\}\right], \tag{4.69}$$

while the explicit form of the probability generating function is given by Theorem 5.1 in Ethier and Griffiths (1990), where ρ is replaced by $4c$, and we have

$$\mathbb{E}[s^{r_n(\infty)}] = \frac{R_n(v)}{R_1(v)}, \tag{4.70}$$

where

$$\begin{aligned} R_n(v) &= \int_0^1 x^{4c(1-v)-1}(1-x)^{n-1}e^{-4cv(1-x)}dx \\ &= \frac{(n-1)!}{(4c(1-v))_n}{}_1F_1(n;4c(1-v)+n;-4cv) \\ &= \sum_{i=0}^{\infty}\frac{(n+i-1)!(-4cv)^i}{(4c(1-v)+n)_{n+i}i!}. \end{aligned}$$

(4.69) provides a way to compute the expectation of the total length of the edges in the ancestral selection graph in the time interval $(0, W^c_{n,1})$, and we have

$$\mathbb{E}\left[\int_0^{W^c_{n,1}} b^1_n(u)du\right] = \frac{1}{2c}\mathbb{E}[r_n(\infty)] = \sum_{k=1}^{\infty}(4c)^{k-1}\sum_{m=1}^{n-1}\frac{1}{(m)_k}. \tag{4.71}$$

It is possible to obtain the probability that the modified process $\{b^1_n(t); W^c_{n,1} \geq t \geq 1\}$ hits the states $n+1, n+2, ...$

THEOREM 4.4.3. *Let* $z(1) = 0, z(2) = 1$, *and*

$$z(j) = 1 + \frac{\alpha_2}{\beta_2} + \frac{\alpha_2\alpha_3}{\alpha_2\beta_3} + \cdots + \frac{\alpha_2\alpha_3\cdots\alpha_{j-1}}{\beta_2\beta_3\cdots\beta_{j-1}} = \sum_{k=0}^{j-2}\frac{k!}{(4c)^k}, \qquad j = 3, 4, ... \tag{4.72}$$

Then, the probability that the modified process $\{b^1_n(t); W^c_{n,1} \geq t \geq 1\}$ *hits the states* $m = n+1, n+2, ...$ *is*

$$\mathbb{P}[W^c_{n,1} > W^c_{n,m}] = \frac{z(n)}{z(m)}. \tag{4.73}$$

PROOF. The theorem follows from standard results on birth and death processes (see, Karlin and Taylor (1975), pp. 323). It is straightforward to show that $z(b^1_n(t))$ is a martingale to the modified process. $\min\{W^c_{n,1}, W^c_{n,m}\}$ is a Markov time with respect to the modified process. We apply the optimal sampling theorem to conclude that

$$z(n) = \mathbb{E}[b^1_n(\min\{W^c_{n,1}, W^c_{n,m}\})] = \mathbb{P}[W^c_{n,1} > W^c_{n,m}]z(m), \qquad m = n+1, n+2, ... \tag{4.74}$$

□

REMARK 4.4.4. *For small* c, $\mathbb{P}[W^c_{n,1} < W^c_{n,m}]$ *can be expanded into a power series in* c.

$$\mathbb{P}[W^c_{n,1} > W^c_{n,m}] = \frac{(4c)^{m-n}}{[m-2]_{m-n}} + O(c^{m-n+1}), \qquad m = n+1, n+2, ... \tag{4.75}$$

4.5. Time to fixation

In studying evolutionary process from the standpoint of population genetics, the probability and the time to fixation of a mutant gene play important roles. The expected time to fixation of a mutant gene conditional on fixation was obtained by Kimura and Ohta (1969). Furthermore, Ewens (1973a) and Maruyama and Kimura (1974) showed that expected length of time which it takes for an allele to increase frequency from q to y $(> q)$ on the way to fixation is equal to the expected length of time which the same allele takes when its frequency decrease from y to q on the way to extinction. The time-reversibility property is equivalent to the property that the density of the expected sojourn time does not depend on the sign of the selection coefficient, which was shown by Maruyama (1972). While these results are well known, their interpretation in terms of the ancestral process of the whole population $\{b_\infty(t); t \geq 0\}$ are interesting.

The fixation probability was obtained by solving the Kolmogorov backward equation for the diffusion process $\{x_p(t); t \geq 0\}$ (Kimura (1957)). The fixation probability of the allele A_1 is

$$u_1(p) = \frac{1 - e^{-4cp}}{1 - e^{-4c}}, \tag{4.76}$$

and the fixation probability of the allele A_2 is $1 - u_1(p)$. It follows from (4.9) that

$$1 - u_1(p) = 2(1 - r^2)e^{c(r-1)} \sum_{k=0}^{\infty} \frac{V_{1k}^{(1)}(c, r)V_{1k}^{(1)}(c, 1)}{2\lambda_k}. \tag{4.77}$$

It is possible to obtain the fixation probability from the stationary measure of the ancestral process (4.33). If the allele A_2 fixes in a population, the ancestral particles of the whole population in infinite time backwards consist of type A_2 particles only, and we have

$$\mathbb{E}[q^{b_\infty(\infty)}] = \sum_{i=1}^{\infty} \pi_i q^i = \frac{e^{4cq} - 1}{e^{4c} - 1} = 1 - u_1(p). \tag{4.78}$$

The density of time to fixation of the allele A_2 conditional on fixation has a genealogical interpretation. Let

$$T_0^c := \inf\{t \geq 0; y_q(t) = 1\}. \tag{4.79}$$

Then, it follows from the expression

$$\mathbb{P}[T_0^c < t | T_0^c < \infty] = \frac{\mathbb{E}[q^{b_\infty(t)}]}{1 - u_1(p)} = \frac{\sum_{i=1}^{\infty} \mathbb{P}[b_\infty(t) = i]q^i}{\sum_{i=1}^{\infty} \pi_i q^i} \tag{4.80}$$

that the process of fixation of the allele in a diffusion model, in which left hand side converges to one as $t \to \infty$, corresponds to convergence of the distribution of the ancestral process $\mathbb{P}[b_\infty(t) = i]$ to its stationary measure π_i as $t \to \infty$.

The expected time to fixation of the allele A_2 conditional on fixation was obtained by solving the Kolmogorov backward equation (Kimura and Ohta (1969), Maruyama (1972)), and

$$\mathbb{E}[T_0^c | T_0^c < \infty] = \int_0^1 \Phi(q, y) dy, \tag{4.81}$$

where $\Phi(q, y)$ is the density of the expected sojourn time of the allele A_2 at frequency y in the path starting from frequency q and going to fixation, and

$$\begin{aligned} \Phi(q,y) &= \frac{S(y)S(1-y)}{2cy(1-y)S(1)}, \qquad y > q, \\ &= \frac{S(y)}{2cy(1-y)} \left\{ \frac{S(1-y)}{S(1)} - \frac{S(q-y)}{S(q)} \right\}, \qquad y < q, \end{aligned} \tag{4.82}$$

and $S(y) = \exp(4cy) - 1$. Then,

$$\mathbb{E}[T_0^c | T_0^c < \infty] = \int_0^1 \frac{S(y)S(1-y)}{2cy(1-y)S(1)} dy - \int_0^q \frac{S(y)S(q-y)}{2cy(1-y)S(q)} dy, \tag{4.83}$$

where

$$\begin{aligned} \int_0^1 \frac{S(y)S(1-y)}{2cy(1-y)S(1)} dy &= \frac{\pi_1}{8c^2} \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \frac{(4c)^{i+j}}{i!j!} \int_0^1 y^{i-1}(1-y)^{j-1} dy \\ &= \frac{\pi_1}{2c} \sum_{k=1}^{\infty} \frac{(4c)^k}{(k+1)!} \sum_{i=1}^{k} \frac{1}{i(k-i+1)} \\ &= 4\pi_1 \sum_{k=0}^{\infty} \frac{H_{k+1}(4c)^k}{(k+2)!}, \end{aligned}$$

$H_k = 1 + 1/2 + \cdots + 1/k$, and

$$\begin{aligned} \int_0^q \frac{S(y)S(q-y)}{2cy(1-y)S(q)} dy &= \frac{1}{2cS(q)} \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \frac{(4c)^{i+j}}{i!j!} \int_0^q \frac{y^{i-1}(q-y)^j}{1-y} dy \\ &= \frac{1}{2cS(q)} \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \frac{(4cq)^{i+j}}{i(i+j)!} {}_2F_1(1, i, i+j+1; q) \\ &= \frac{1}{2cS(q)} \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \frac{(4cq)^{i+j}}{i(i+j)!} \sum_{k=0}^{\infty} \frac{(i)_k q^k}{(i+j+1)_k}. \end{aligned}$$

It is possible to obtain the expected time to fixation of the allele A_2 conditional on fixation from the distribution $b_\infty(t)$, and we have

$$\begin{aligned}\mathbb{E}[T_0^c|T_0^c<\infty] &= \frac{1}{\sum_{i=1}^\infty \pi_i q^i}\int_0^\infty t\frac{d}{dt}\left[\sum_{i=1}^\infty\{\mathbb{P}[b_\infty(t)=i]-\pi_i\}q^i\right]dt \\ &= \frac{1}{\sum_{i=1}^\infty \pi_i q^i}\sum_{i=1}^\infty q^i\int_0^\infty\{\mathbb{P}[b_\infty(t)=i]-\pi_i\}dt. \qquad (4.84)\end{aligned}$$

From the two expressions (4.83) and (4.84), an identity at $q=0$ follows immediately.

$$\sum_{k=0}^\infty \frac{V_{1k}^{(1)}(c,-1)V_{1k}^{(1)}(c,1)}{\lambda_k^2 N_{1k}} = e^{2c}\pi_1^2\sum_{k=0}^\infty\frac{H_{k+1}(4c)^k}{(k+2)!}. \qquad (4.85)$$

It is straightforward to obtain similar identities by comparing (4.83) and (4.84) in each power of q. Moreover, explicit form of the higher moments of the time to fixation conditional on fixation (Maruyama (1972)) is available, and they produce similar identities.

The density of time to fixation of a single mutant gene conditional on fixation has interesting properties. Let

$$T_1^c := \inf\{t\geq 0; x_p(t)=1\}. \qquad (4.86)$$

Then, from a time-reversibility argument on the conditional diffusion process (Ewens (1973a), Maruyama and Kimura (1974)), we have

$$\lim_{q\to 0}\mathbb{P}[T_0^c<t|T_0^c<\infty] = \lim_{p\to 0}\mathbb{P}[T_1^c<t|T_1^c<\infty] = \frac{\mathbb{P}[b_\infty(t)=1]}{\pi_1}. \qquad (4.87)$$

The same density hold for a mutant gene of allele A_1 and a mutant gene of allele A_2. This property has an intuitive genealogical interpretation. The conditional density is given by the probability of the whole population being descended from a single real ancestral particle. Since there is no variation in the population, selection cannot have an effect on it and consequently, the conditional density should not depend on the allelic type. (4.4), (4.39) and Corollary 4.3.4 gives bounds for the density of time to fixation of a single mutant gene conditional on the fixation, and

$$\frac{1}{\pi_1}-\frac{3e^{-c}\rho_2^c(t)+2c}{\pi_1\lambda_0}\leq\lim_{q\to 0}\mathbb{P}[T_0^c<t|T_0^c<\infty]\leq\frac{1}{\pi_1}-\max\left\{\frac{\rho_2^0(t)}{\pi_1},\frac{3e^{-c}\rho_2^c(t)+2c}{\pi_1(3e^{-c}+2c)}\right\}. \qquad (4.88)$$

It is worth noting that the identity (4.87) gives following identity in the distribution $b_n(t)$. Its interpretation in terms of the ancestral process $\{b_n(t);t\geq 0\}$ is unclear.

REMARK 4.5.1. *The time-reversibility argument on the conditional diffusion process gives*

$$\mathbb{P}[b_\infty(t)=1] = \lim_{n\to\infty} e^{-4c}\sum_{k=1}^{n}\frac{(-1)^{k+1}n!}{k!(n-k)!}\mathbb{E}[b_k(t)]. \tag{4.89}$$

PROOF. (4.87) is equivalent to

$$\lim_{p\to 0}\frac{f(p,1;t)}{u_1(p)} = \lim_{q\to 0}\frac{f(p,0;t)}{1-u_1(p)} = \frac{\mathbb{P}[b_\infty(t)=1]}{\pi_1}, \tag{4.90}$$

where

$$\begin{aligned}\lim_{p\to 0}\frac{f(p,1;t)}{u_1(p)} &= \lim_{p\to 0}\lim_{n\to\infty}\frac{\mathbb{E}[x_p(t)^n]}{u_1(p)} = \lim_{p\to 0}\lim_{n\to\infty}\frac{\mathbb{E}[(1-y_q(t))^n]}{u_1(p)}\\ &= \lim_{p\to 0}\lim_{n\to\infty}\sum_{k=0}^{n}\frac{(-1)^k n!}{k!(n-k)!}\frac{\mathbb{E}[q^{b_k(t)}]}{u_1(p)} = \lim_{n\to\infty} e^{-4c}\sum_{k=1}^{n}\frac{(-1)^{k+1}n!}{k!(n-k)!}\frac{\mathbb{E}[b_k(t)]}{\pi_1}. \end{aligned} \tag{4.91}$$

□

In the neutral diffusion process, the density of time to fixation of a mutant gene conditional on fixation follows

$$\lim_{q\to 0}\mathbb{P}[T_0^0 < t | T_0^0 < \infty] = \mathbb{P}[a_\infty(t)=1], \tag{4.92}$$

where T_0^0 is the time to fixation of a mutant gene in the neutral diffusion process. From (4.83), the expected time to fixation of a mutant gene conditional on fixation has a simple form

$$\lim_{q\to 0}\mathbb{E}[T_0^c | T_0^c<\infty] = 4\pi_1\sum_{j=0}^{\infty}\frac{H_{j+1}(4c)^j}{(j+2)!} < \lim_{q\to 0}\mathbb{E}[T_0^0|T_0^0<\infty] = 2, \tag{4.93}$$

where the inequality holds from the following lemma:

LEMMA 4.5.2. *The density of expected sojourn time of the allele A_2 at frequency y in the path starting from frequency 0 and going to fixation satisfies*

$$\frac{S(y)S(1-y)}{2cy(1-y)S(1)} < 2, \qquad 0<y<1. \tag{4.94}$$

PROOF. The inequality is equivalent to

$$\frac{e^{4cy}-1}{y}\frac{e^{4c(1-y)}-1}{1-y} < 4c(e^{4c}-1), \tag{4.95}$$

or

$$\sum_{i=0}^{\infty}\sum_{j=0}^{i}\frac{(4c)^i y^j(1-y)^{i-j}}{(j+1)!(i-j+1)!} < \sum_{i=0}^{\infty}\frac{(4c)^i}{(i+1)!}. \tag{4.96}$$

The inequality follows from an inequality

$$\sum_{j=0}^{i} \frac{y^j(1-y)^{i-j}}{(j+1)!(i-j+1)!} < \frac{1}{(i+1)!}\sum_{j=0}^{i} \frac{i!y^j(1-y)^{i-j}}{j!(i-j)!} = \frac{1}{(i+1)!}, \qquad i = 0, 1, ... \tag{4.97}$$

□

As c becomes large, $\mathbb{P}[b_\infty(t) = 1]$ decreases, while the expected fixation time of a mutant gene conditional on fixation decreases. It is straightforward to show that the inequality for the expected fixation time (4.93) is equivalent to an inequality

$$\int_0^\infty \left\{ \frac{\mathbb{P}[b_\infty(t) = 1]}{\pi_1} - \mathbb{P}[a_\infty(t) = 1] \right\} dt > 0. \tag{4.98}$$

In the neutral process, the density of the waiting time to the ancestral process hits the state 1 and the conditional fixation time are given by the probability that the ancestral process is at the state 1 (4.3,4.92). It follows that

$$\mathbb{E}[W^0_{\infty,1}] = \lim_{q\to 0} \mathbb{E}[T^0_0 | T^0_0 < \infty] = \int_0^\infty \{\mathbb{P}[a_\infty(t) = 1] - 1\} dt = 2. \tag{4.99}$$

In contrast, in the process with selection, we have

$$\mathbb{E}[W^c_{\infty,1}] = 2\sum_{j=0}^{\infty} \frac{(4c)^j}{(j+1)(j+1)!} > 2, \tag{4.100}$$

while

$$\lim_{q\to 0} \mathbb{E}[T^c_0 | T^c_0 < \infty] = \int_0^\infty \left\{ \frac{\mathbb{P}[b_\infty(t) = 1]}{\pi_1} - 1 \right\} dt = 4\pi_1 \sum_{j=0}^{\infty} \frac{H_{j+1}(4c)^j}{(j+2)!} < 2. \tag{4.101}$$

4.6. Summary

In this article, properties of the ancestral process $\{b_n(t); t \geq 0\}$, the total number of the real and the virtual particles, were investigated. An explicit form of the probability distribution of $b_n(t)$ was obtained. Altohugh this expression cannot be given in closed form, since it involves eigenvalues and coefficients which are determined by an intractable three-term recursion relation, it is possible to expand the probability distribution as a perturbation series in $2c$. This expression is given in closed form for each order of the perturbation and is accurate when $|c|$ is small.

If a sample consists only of type A_2 particles, the probability distribution of the ancestral particles, all of which are A_2, is $b_n(t)$ (see Theorem 1). If a sample contains type A_1 particles, the joint probability distribution of the number of the A_1 particles and the number of the A_2 particles is interesting. However, it seems that the expression of the moments in the diffusion model (4.17) does not give any insights of the joint probability

distribution of the ancestral particles, except for the case that a sample consists of a single A_1 particle. The joint distribution of the ancestral particles needs further investigations.

The time-reversibility argument of the conditional diffusion process gives an identity, whose interpretation in terms of the ancestral process is unclear (see Remark 5.1). The interpretation of the time-reversibility in terms of the ancestral process needs further investigations.

4.7. Appendix. The oblate spheroidal wave function

The oblate spheroidal wave function $V_{1k}^{(1)}(c,z)$ can be represented by expansions of the form (Stratton et al. (1941))

$$V_{1k}^{(1)}(c,z) = {\sum_{l\geq 0}}' f_l^k(c) T_l^1(z), \qquad k=0,1,... \tag{4.102}$$

This notation was used in Kimura (1955c). It was denoted by $V_{1k}^{(1)}(-ic,z)$ in Stratton et al. (1941) and $(1-z^2)^{\frac{1}{2}}S_{1k+1}(c,z)$ in Flammer (1957). From the orthogonal properties of the Gegenbauer function it is shown that

$$\int_{-1}^{1}(1-z^2)V_{1k}^{(1)}(c,z)V_{1l}^{(1)}(c,z)dz = \delta_{k,l}N_{1k}, \tag{4.103}$$

where

$$N_{1k} = 2{\sum_{l\geq 0}}' \frac{(l+1)(l+2)}{(2l+3)}(f_l^k(c))^2.$$

Note that

$$V_{1k}^{(1)}(c,1) = \frac{1}{2}{\sum_{l\geq 0}}'(l+1)(l+2)f_l^k(c), \qquad V_{1k}^{(1)}(c,-1) = (-1)^k V_{1k}^{(1)}(c,1). \tag{4.104}$$

The coefficients $f_l^k(c)$ satisfy a three-term recursion in the form

$$A_{l+2}f_{l+2}^k(c) + B_l f_l^k(c) + C_{l-2}f_{l-2}^k(c) = 0, \tag{4.105}$$

where

$$A_l = -\frac{(l+1)(l+2)}{(2l+1)(2l+3)}, \qquad B_l = \frac{l(l+3)-b_k}{c^2} - \frac{2l^2+6l+1}{(2l+1)(2l+5)}, \qquad C_l = -\frac{(l+1)(l+2)}{(2l+3)(2l+5)},$$

and $b_k = 2\lambda_k - 2 - c^2$. $f_l^k(c) = 0$ for odd l if k is even and for even l if k is odd. (4.105) can be developed as a continued fraction.

$$\begin{aligned} \frac{f_l^k}{f_{l+2}^k} &= -\frac{A_{l+2}}{B_l-}\frac{C_{l-2}A_l}{B_{l-2}-}\cdots\frac{C_2A_4}{B_2-}\frac{A_2}{B_0} \qquad l=0,2,... \\ &\quad -\frac{A_{l+2}}{B_l-}\frac{C_{l-2}A_l}{B_{l-2}-}\cdots\frac{C_3A_5}{B_3-}\frac{A_3}{B_1} \qquad l=1,3,... \end{aligned} \tag{4.106}$$

and

$$\frac{f_{l+2}^k}{f_l^k} = -\frac{C_l}{B_{l+2}-}\frac{A_{l+4}C_{l+2}}{B_{l+4}-}\cdots, \qquad l = 0, 1, .. \tag{4.107}$$

b_k is determined by the condition that the reciprocal of the ratio f_l/f_{l+2} by (4.106) must equal the value of f_{l+2}/f_l obtained from (4.107). Then, the continued fractions provide a way to compute arbitrary coefficient.

For small c, the eigenvalue can be expanded into a power series in c.

$$\lambda_k = \frac{(k+1)(k+2)}{2} + \frac{(k+1)(k+2)}{(2k+1)(2k+5)}c^2 + O(c^4). \tag{4.108}$$

If we set $f_k^k(c) = 1$, then

(4.109)

$$f_{k+2}^k(c) = \frac{(k+1)(k+2)}{2(2k+3)(2k+5)^2}c^2 + O(c^4), \qquad f_{k-2}^k(c) = -\frac{(k+1)(k+2)}{2(2k+1)^2(2k+3)}c^2 + O(c^4),$$

and other coefficients are zero up to $O(c^4)$.

For large c, an asymptotic expansion is possible (Flammer (1957)), and

(4.110)

$$V_{1k}^{(1)}(c,z) \simeq \sum_{i=-\nu_k}^{\infty} A_i^{1k+1}\left\{e^{-c(1-z)}L_{\nu_k+i}^{(1)}[2c(1-z)] + (-1)^k e^{-c(1+z)}L_{\nu_k+i}^{(1)}[2c(1+z)]\right\},$$

where $\nu_k = k/2$ for k even and $\nu_k = (k-1)/2$ for k and $L_{\nu_k+i}^{(1)}(\cdot)$ is the Laguerre functions. The coefficients A_i^{1k+1} are given in Flammer (1957). The eigenvalue is

$$\lambda_k = 2(1+\nu_k)c - \nu_k(\nu_k+2) - 1 + O(c^{-1}). \tag{4.111}$$

Since ν_k is the same when k is equal to k' and when k is equal to $k'+1$, where k' is an even integer, pair of eigenvalues coalesce as c becomes large.

CHAPTER 5

Selective sweep

5.1. Introduction

DNA sequence data are a rich source for detecting adaptive evolution. Especially, huge single nucleotide polymorphism (SNP) data with linkage phase, or SNP haplotype data, are emerging. Developing powerful statistical methods to detect positive selection with the haplotype data is an important issue in population genetics. The Statistical methods using within species polymorphism data can be loosely classified into three categories: site frequency, haplotype frequency, and linkage disequilibrium methods. The site frequency methods require only frequencies of variants at polymorphic nucleotide sites. Linkage phase of these variants is neither required nor used. One sub-category of the methods is based on the completely linked infinite site model and utilize the simple summary statistics of site frequency spectrum (e.g. Tajima (1989a); Fu and Li (1993); Fay and Wu (2000)). The other sub-category of the methods is based on single site models and utilize the site frequency spectrum at unlinked segregating sites (e.g., Kim and Stephan (2002); Nielsen et al. (2006)). The haplotype frequency methods require additional information on the linkage phase among variant sites. A haplotype is scored as an allele and conditional haplotype frequency spectrum are used for detection. One sub-category of the method is based on the infinite allele model and utilize allele frequency spectrum conditional on the number of different haplotypes (Ewens, 1973b; Watterson, 1978; Slatkin, 1994b). The other sub-category of the methods is based on the infinite sites model and utilize allele frequency spectrum conditional on the number of segregating sites (Depaulis and Veuille, 1998; Innan et al., 2005). Recombination is not considered in these null distribution. It can be expected that there are few intra-haplotype recombination, since a long range haplotype which has experienced many recombinations makes no biological sense, nevertheless, the impact of recombination to the power and robustness of these tests should be addressed in statistical point of view. The last category, the linkage disequilibrium methods, in which the most popular one is probably the extended haplotype homozygosity test (Sabeti et al., 2002), depends heavily on simulations and attracts little theoretical interests.

Recently, the author and coworkers assessed the power and robustness of the haplotype and site frequency methods to detect positive selection by extensive simulations (Zeng et al., 2007). In their study, intra haplotype recombination are incorporated. They found that although the haplotype frequency methods conditional on the number of haplotypes were developed for nonrecombining haplotypes, these tests are insensitive to intra-haplotype recombination. It could be said that the number of haplotypes have information of not only mutation rate but also recombination rate. Such tests can therefore be applied to recombining haplotypes. In contrast, tests conditional on the number of segregating sites become overly conservative in the presence of recombination. In fact, the Watterson's homozygosity test (Watterson, 1978) was usually the most powerful test during the sweep phase, especially when the local recombination rate is high. The extended haplotype homozygosity test relies heavily on the prior knowledge of the target of selection. With the knowledge, it is the most powerful test, whereas in the absence of this prior information, the test has little power.

Although the tests based on summary statistics of the allele frequency spectrum conditional on the number of haplotypes are generally powerful, these tests give no insights how the selection operates. On the other hand, the site frequency methods utilize the site frequency spectrum at unlinked segregating sites are likelihood based and provide maximum likelihood estimates of the position of the target of selection and the selection intensity. In theoretical and practical points of view, however, there are ambiguities in their use of composite likelihood among weakly linked site. The composite likelihood is not a real likelihood, the test should be done by using simulated distribution empirically. It could be ambiguous which site among the weekly linked sites contribute mainly to the results. If the segregating sites were strictly unlinked, we could avoid the ambiguity. However, we could not expect gain of power by pooling unlinked sites. Their likelihood is based on single site bi-allelic models, the information could be used at each site would be small and the power therefore would be daunting. In this chapter, a new likelihood based test to detect a recent sweep is presented, which is a natural extension of the tests based on summary statistics of the allele frequency spectrum conditional on the number of haplotypes. For the likelihood for a model at the end of the sweep, a sampling formula is employed, which was presented by the author (Mano, 2006).

5.2. Sampling distribution at the end of a selective sweep

Let us assume a neutral locus A in which alleles are segregating with frequency distribution which follows that of the infinite allele model. In this subsection recombination are ignored and a haplotype is scored as an allele. Let $C_j,\ j = 1, 2, ..., n$ be the number of different types of alleles whose frequency in a sample of size n is j. Under neutrality, the distribution of $\boldsymbol{C}$ is given by the Ewens sampling formula (Ewens, 1972)

$$Q_n(\boldsymbol{c}) = I(|\boldsymbol{c}| = n)\frac{n!}{(\theta)_n}\prod_{j=1}^{n}\left(\frac{\theta}{j}\right)^{c_j}\frac{1}{c_j!}, \tag{5.1}$$

where $|\boldsymbol{c}| := \sum_{j=1}^{n} jc_j$, $\theta = 4Nu$ and u is the mutation rate per generation. Conditioned by the number of different types of alleles, we have

$$\mathbb{P}[\boldsymbol{C} = \boldsymbol{c}|\|\boldsymbol{C}\| = k] = \frac{n!}{\mathfrak{S}_n^k}\prod_{j=1}^{n}\frac{1}{j^{c_j}c_j!}, \tag{5.2}$$

where $\|\boldsymbol{c}\| := \sum_{j=1}^{n} c_j$ and $\mathfrak{S}_n^k$ is the unsigned Stirling number of the first kind. It is remarkable that the conditional distribution does not depend on θ. The number of types of alleles is the sufficient statistic of θ. The Watterson's homozygosity test (Watterson, 1978) is a test by using the homozygosity $\sum_{j=1}^{n} c_j^2$ as a test statistic to reject neutrality.

Suppose the neutral locus A links to a locus B in which an advantageous mutant appears. Assume the selective advantage of the mutant allele B over the wildtype allele b is s and the genic selection is assumed. We are interested in how the distribution of $\boldsymbol{C}$ in the locus A could be affected by linking to the locus B. Using a deterministic model, Maynard Smith and Haigh (1974) obtained the frequency of the hitchhikers, or the descendants of the allele at which was carried at the locus A of the founder chromosome at the end of the sweep. That is (Eq. 14 of Maynard Smith and Haigh (1974))

$$p_M \approx \frac{r}{s}\log 2N, \tag{5.3}$$

where r is the recombination fraction between the two loci. Apart form hitchhikers, there are also a few free-riders who jump on the sweep when it is already under way, and make it as singletons into the sample. The number of the hitchhikers follows $Binom.(n, p_M)$. To include the randomness of the frequency path, especially at the beginning of the sweep, Etheridge et al. (2006) investigated approximation of a genealogy of the alleles in the locus A in the random background of the alleles in the locus B by the structured coalescent. If the allele B is destined to be fix, the random conditional path $\{x_p(t); T_1 \geq t \geq 0\}$, where

$T_1 = \inf\{t \geq 0; x_p(t) = 1\}$, follows a stochastic differential equation

$$dx = \sqrt{x(1-x)}dB + \alpha \coth(\alpha x)x(1-x)dt, \tag{5.4}$$

where $\alpha = 2Ns$ and time is measured in unit of $2N$ generations (Ewens, 1973a). Etheridge et al. (2006) showed that the probability that a single ancestral lineage is not hit by a recombination is approximately given by $p = e^{-\gamma}$, where

$$\gamma = \rho \frac{\log \alpha}{\alpha}, \tag{5.5}$$

and $\rho = 2Nr$. For large α the subpopulation carrying the advantageous allele is expanding quickly near the beginning of the sweep, the sample genealogy in the B background can be approximated by a star-shaped genealogy, i.e. lineages all coalescing at the beginning of the sweep. In addition, Etheridge et al. (2006) showed that the probability that a neutral lineage recombines out of the B background and then recombines back into the B background and that a pair of neutral lineages coalesces in the b background are $O\{(\log \alpha)^{-2}\}$. Therefore, with an error $O\{(\log \alpha)^{-2}\}$, we may ignore back-recombinations. A first approximation to the number of hitchhikers, or nonrecombinants, is given by $Binom.(n, p)$-number of individuals stemming from the founder of the sweep, and the rest being free-riders, or recombinants, all having different ancestors at the beginning of the sweep (See Figure 5.1). Etheridge et al. (2006) gave a proof that the sampling distribution is accurate with probability $1 - O(1/\log \alpha)$. Consequently, (5.3) is accurate with the same probability.

Let us consider the sampling distribution at the neutral locus A at the end of a sweep, where we assume that the distribution at the beginning of the sweep followed the Ewens sampling formula and the distribution of recombinants and non-recombinants after the sweep follows the Etheridge et al. (2006) sampling distribution discussed above. Figure 5.1 shows an example of a genealogy under a selective sweep. Lineages in the allele b background are dotted. Colors represents allelic types in the locus A. Lineages connected at the beginning of the sweep are the same allelic type at the beginning. Here, the distribution is $c_1 = 2, c_2 = 1, c_6 = 1$. The number of the recombinant is 6 and the number of non-recombinant is 4. Note that the allelic type frequency of the founder of the sweep is not the same as the number of the non-recombinant, since there could be recombinants whose allelic type is the same as the founder. Let l be the allelic type frequency of the founder of the sweep. The distribution after the sweep is given by a weighted binomial mixture of the Ewens distribution, where the summations are taken

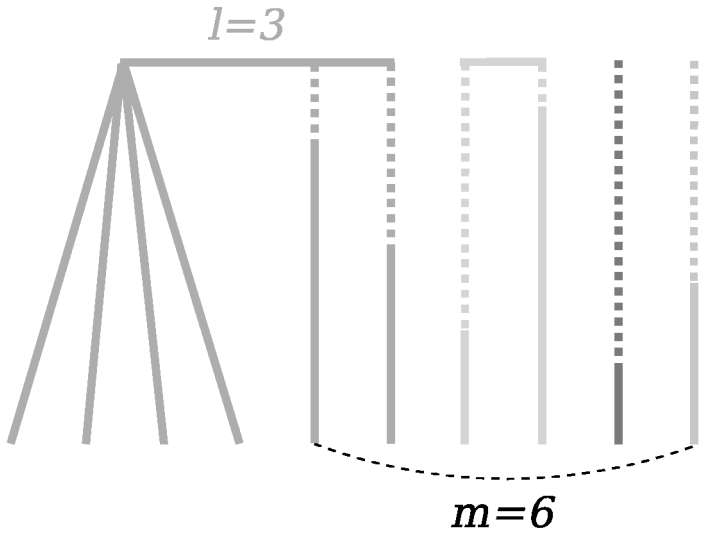

FIGURE 5.1. A genealogy of ten samples under a selective sweep.

for all possible configurations characterized by the number of the recombinants $m, m = 0, 1, ..., n)$ and $l, l = 1, 2, ..., m + 1$.

THEOREM 5.2.1. *For* $n = 3, 4, ...$

$$\begin{aligned}
\tilde{Q}_n(\boldsymbol{c}) &= Q_n(\boldsymbol{c})[(1-p)^n + n(1-p)^{n-1}p] \\
&\quad + \sum_{m=\lceil (n-1)/2 \rceil}^{n-2} \left\{ \sum_{l=1}^{2m-n+2} \frac{l(c_l+1)[n]_m}{(m+1)!}(1-p)^m p^{n-m} Q_{m+1}(\boldsymbol{c}+\boldsymbol{e}_l-\boldsymbol{e}_{n-m+l-1}) \right. \\
&\quad \left. + \sum_{l=2m-n+3}^{m-\|\boldsymbol{c}\|+2} \frac{l(c_l+1)[n]_m}{(m+1)!}(1-p)^m p^{n-m} I(c_{n-m+l-1}=1) Q_{m+1}(\boldsymbol{c}+\boldsymbol{e}_l) \right\} \\
&\quad + \sum_{m=\|\boldsymbol{c}\|-1}^{\lceil (n-3)/2 \rceil} \sum_{l=1}^{m-\|\boldsymbol{c}\|+2} \frac{l(c_l+1)[n]_m}{(m+1)!}(1-p)^m p^{n-m} I(c_{n-m+l-1}=1) Q_{m+1}(\boldsymbol{c}+\boldsymbol{e}_l). \qquad (5.6)
\end{aligned}$$

PROOF. The first term is trivial to obtain. $1 \leq l \leq m - \|\boldsymbol{c}\| + 2$ because the number of recombinants whose allelic type is other than the founder should be equal to or larger than $\|\boldsymbol{c}\| - 1$. If $m + 1 < n - m + l - 1$, then $I(c_{n-m+l-1} = k),\ k = 1, 2, ...$ is needed, since $Q_{m+1}(\cdot)$ do not account alleles whose frequency is $n - m + l - 1$. The inequality is equivalent to $m < (n + l)/2 - 1$. Assume $m \leq \lceil (n-3)/2 \rceil$. If $c_{n-m+l-1} = k$, then $k(n - m + l - 1) \leq n$. The inequality is equivalent to $m \geq n(k-1)/2 + l - 1$, which holds if and only if $k = 1$. Next, assume $m \geq \lceil (n-1)/2 \rceil$. If $l \leq 2m - n + 2$, then $Q_{m+1}(\cdot)$ capture alleles whose frequency is $n - m + l - 1$. Otherwise, assume $c_{n-m+l-1} = k$, then

$2m-n+3 \le l \le m+1-n(k-1)/2$. Such l exists if and only if $k=1$. Summing probabilities for these possibilities, the theorem follows. □

For small γ,

$$\begin{aligned}\tilde{Q}_n(\boldsymbol{c}) &= \sum_{m=\|c\|-1}^{\lceil (n-3)/2\rceil}\sum_{l=1}^{m-\|\boldsymbol{c}\|+2}\frac{l(c_l+1)[n]_m}{(m+1)!}(1-p)^m p^{n-m} I(c_{n-m+l-1}=1)Q_{m+1}(\boldsymbol{c}+\boldsymbol{e}_l)\\ &\quad +O(\gamma^{(n-1)/2})\\ &= \sum_{m=\|\boldsymbol{c}\|-1}^{\lceil (n-3)/2\rceil}\sum_{l=1}^{m-\|\boldsymbol{c}\|+2}\frac{[n]_m\theta^{\|\boldsymbol{c}\|}}{(\theta)_{m+1}}(1-p)^m p^{n-m}\prod_{j=1}^{m-l+1}\frac{1}{j^{c_j}c_j!}\\ &\quad \times I(c_{n-m+l-1}=1,|\boldsymbol{c}'|=m-l+1)+O(\gamma^{(n-1)/2}), \qquad (5.7)\end{aligned}$$

where $\boldsymbol{c}'$ is a $m+1$-dimensional vector, where $c'_j = c_j$ for $j=1,2,...,m+1$. It is remarkable that if $\|\boldsymbol{c}\| > \lceil (n-1)/2\rceil$ then the probability that $\boldsymbol{c}$ is a result of a recent sweep is $O(\gamma^{(n-1)/2})$. It is straightforward to obtain the distribution of the number of types of alleles:

$$\begin{aligned}\mathbb{P}[\|\boldsymbol{C}\|=k] &= \sum_{\|\boldsymbol{c}\|=k}\tilde{Q}_n(\boldsymbol{c})\\ &= \sum_{m=k-1}^{\lceil (n-3)/2\rceil}\sum_{l=1}^{m-k+2}\frac{[n]_m\theta^k}{(\theta)_{m+1}}(1-p)^m p^{n-m}\\ &\quad \times\sum_{\|\boldsymbol{c}'\|=k-1}\prod_{j=1}^{m-l+1}\frac{1}{j^{c_j}c_j!}I(|\boldsymbol{c}'|=m-l+1)+O(\gamma^{(n-1)/2})\\ &= \sum_{m=k-1}^{\lceil (n-3)/2\rceil}\sum_{l=1}^{m-k+2}\frac{[n]_m\theta^k\mathfrak{S}^{k-1}_{m-l+1}}{(\theta)_{m+1}(m-l+1)!}(1-p)^m p^{n-m}+O(\gamma^{(n-1)/2}). \qquad (5.8)\end{aligned}$$

Then, the conditional distribution is

$$\mathbb{P}[\boldsymbol{C}=\boldsymbol{c}|\|\boldsymbol{C}\|=k]=\frac{\tilde{Q}_n(\boldsymbol{c},\|\boldsymbol{c}\|=k)}{\mathbb{P}[\|\boldsymbol{C}\|=k]}. \qquad (5.9)$$

Up to the first order in γ, we have

$$\mathbb{P}[C_{n-k+1}=1,C_1=k-1|\|\boldsymbol{C}\|=k]=1-\frac{(k-1)(n-k+1)}{2(\theta+k)}\gamma+O(\gamma^2), \qquad (5.10)$$

$$\mathbb{P}[C_{n-k}=1,C_2=1,C_1=k-2|\|\boldsymbol{C}\|=k]=\frac{(k-1)(n-k+1)}{2(\theta+k)}\gamma+O(\gamma^2), \qquad (5.11)$$

and the probabilities of other configurations vanish. It is straightforward to obtain the conditional expectation of the number of types of allele $C_i, i=1,2,...,n$, and we have

$$\mathbb{E}[C_i|\|\boldsymbol{C}\|=k]=\frac{\sum_{\|\boldsymbol{c}\|=k}c_i\tilde{Q}_n(\boldsymbol{c},\|\boldsymbol{c}\|=k)}{\mathbb{P}[\|\boldsymbol{C}\|=k]}, \qquad (5.12)$$

where

$$\begin{aligned}\sum_{\|\boldsymbol{c}\|=k} c_i \tilde{Q}_n(\boldsymbol{c}, \|\boldsymbol{c}\| = k) &= \sum_{m=k-1}^{\lceil (n-3)/2 \rceil} \sum_{l=1}^{m-k+2} \frac{[n]_m \theta^k}{(\theta)_{m+1}} (1-p)^m p^{n-m} \\ &\quad \times \sum_{\|\boldsymbol{c}'\|=k-1} \left\{ \prod_{j=1}^{m-l+1} \frac{c_i}{j^{c_j} c_j!} I(|\boldsymbol{c}'| = m-l+1, i \leq m-l+1) \right. \\ &\quad \left. + \prod_{j=1}^{m-l+1} \frac{1}{j^{c_j} c_j!} I(|\boldsymbol{c}'| = m-l+1, i = n-m+l-1) \right\} \\ &\quad + O(\gamma^{(n-1)/2}). \end{aligned} \tag{5.13}$$

For $i = 1, 2, ..., n-k+1$,

$$\begin{aligned}\sum_{\|\boldsymbol{c}\|=k} c_i \tilde{Q}_n(\boldsymbol{c}, \|\boldsymbol{c}\| = k) &= \sum_{m=k+i-2}^{\lceil (n-3)/2 \rceil} \sum_{l=1}^{m-k-i+3} \frac{[n]_m \theta^k \mathfrak{S}_{m-l-i+1}^{k-2}}{i(\theta)_{m+1}(m-l-i+1)!} (1-p)^m p^{n-m} \\ &\quad + \sum_{m=n-i}^{\lceil (n-3)/2 \rceil} \frac{[n]_m \theta^k \mathfrak{S}_{n-i}^{k-1}}{(\theta)_{m+1}(n-i)!} (1-p)^m p^{n-m} + O(\gamma^{(n-1)/2}), \end{aligned} \tag{5.14}$$

and for $i = n-k+2, n-k+3, ..., n$,

$$\begin{aligned}\sum_{\|\boldsymbol{c}\|=k} c_i \tilde{Q}_n(\boldsymbol{c}, \|\boldsymbol{c}\| = k) &= \sum_{m=k+i-2}^{\lceil (n-3)/2 \rceil} \sum_{l=1}^{m-k-i+3} \frac{[n]_m \theta^k \mathfrak{S}_{m-l-i+1}^{k-2}}{i(\theta)_{m+1}(m-l-i+1)!} (1-p)^m p^{n-m} \\ &\quad + O(\gamma^{(n-1)/2}), \end{aligned} \tag{5.15}$$

where $k \neq 1$. Up to the first order in γ, we have

$$\mathbb{E}[C_1 | \|\boldsymbol{C}\|] = (k-1)\left(1 - \frac{n-k+1}{2(\theta+k)}\gamma\right), \tag{5.16}$$

$$\mathbb{E}[C_2 | \|\boldsymbol{C}\|] = \frac{(k-1)(n-k+1)}{2(\theta+k)}\gamma, \tag{5.17}$$

$$\mathbb{E}[C_{n-k} | \|\boldsymbol{C}\|] = \frac{(k-1)(i+1)}{2(\theta+k)}\gamma, \tag{5.18}$$

$$\mathbb{E}[C_{n-k+1} | \|\boldsymbol{C}\|] = 1 - \frac{(k-1)i}{2(\theta+k)}\gamma. \tag{5.19}$$

The conditional expectation of the number of types of alleles for a sample with $n = 50$ is illustrated in Figure 5.2. It is assumed that $\theta = 1, k = 5$, and $\gamma = 0.033$.

5.3. Tests

Various tests have been proposed based on the allele frequency distribution conditional on the number of alleles (5.2). Ewens (1973a) proposed a test whose summary statistics is the frequency of the most common allele, Watterson (1978) proposed a test whose

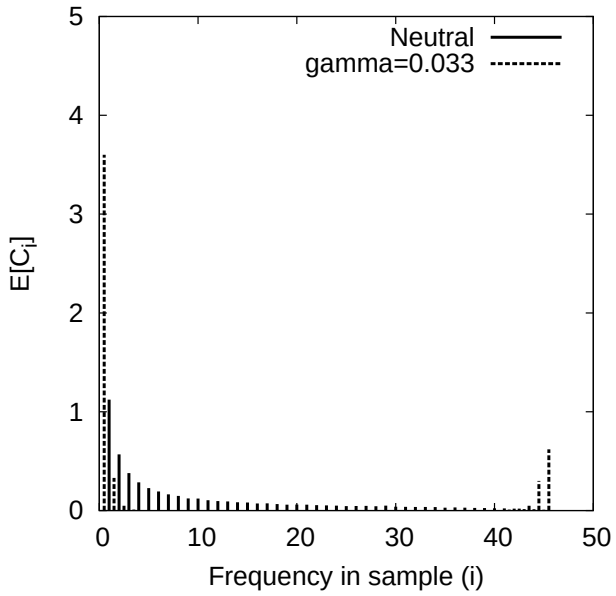

FIGURE 5.2. The conditional expectation of the number of types of alleles whose frequency in a sample is given. $n = 50$, $\theta = 1$ and $k = 5$. The neutral expectation is also shown.

summary statistics is is the homozygosity, and Slatkin (1994b) proposed an exact test based on configuration of haplotype. It was shown that these tests yield similar statistical properties to detect a recent selective sweep (Zeng et al., 2007). In fact, it is straightforward to show that the test proposed by Ewens (1973b) test and that proposed by Watterson (1978) have exactly same power when $\|c\| = 2$. Here, we study the test proposed by Ewens (1973b), because the null distribution of the test statistics is given by a simple analytical form.

LEMMA 5.3.1. *Define a generalized unsigned Stirling number of the first kind* $\mathfrak{S}^k_{n,m}$*, which is the number of permutations of* n *whose decomposition has exactly* k *cycles of which are of length less than* m $(\leq n)$*. Then,* $\mathfrak{S}^k_{n,m}$ *satisfy a recurrence relation*

$$\mathfrak{S}^k_{n,m} = \mathfrak{S}^{k-1}_{n-1,m} + \sum_{i=2}^{\min\{m,n\}} [n-1]_{i-1}\mathfrak{S}^{k-1}_{n-i,m}, \qquad k = 1, 2, ..., n, \tag{5.20}$$

with the boundary conditions $\mathfrak{S}^0_{n,m} = \delta_{n,m}$ *for* $m > n$ *and* $\mathfrak{S}^0_{n,m} = 0$ *for* $m \leq n$*.*

PROOF. Suppose k cycles consist of a cycle which includes n and other $k-1$ cycles. Then, the length of the cycle which include n could be $1, 2, ..., m$. When the length of the cycle is i, $i = 1, 2, ..., m$, the number of ways to choose member of the cycle is ${}_{n-1}C_{i-1}$, and the number of permutations of a cycle whose length is i is $\mathfrak{S}^1_i = (i-1)!$. Then,

contribution of the case to $\mathfrak{S}^k_{n,m}$ is $[n-1]_{i-1}\mathfrak{S}^{k-1}_{n-i,m}$. By summing over i, the lemma follows. □

REMARK 5.3.2. *The number of permutations of n whose decomposition has exactly k cycles of which are of length two or greater is defined as the associated Stirling number of the first kind (Comtet, 1974).*

THEOREM 5.3.3. *The distribution of the frequency of the most common allele G is given by*

$$\mathbb{P}[G \geq g] = 1 - \frac{\mathfrak{S}^k_{n,g-1}}{\mathfrak{S}^k_n}, \qquad g = 1, 2, ..., n. \tag{5.21}$$

PROOF. By using the conditional distribution (5.2) and Lemma 5.3.1, we obtain

$$\begin{aligned} \mathbb{P}[G < g | \|\boldsymbol{C}\| = k] &= \mathbb{P}[C_g = C_{g+1} = \cdots = C_n = 0 | \|\boldsymbol{C}\| = k] \\ &= \sum_{\|\boldsymbol{c}\|=k, c_g=\cdots=c_n=0} \frac{n!}{\mathfrak{S}^k_n} \prod_{j=1}^{n} \frac{1}{j^{c_j} c_j!} = \frac{\mathfrak{S}^k_{n,g-1}}{\mathfrak{S}^k_n}. \end{aligned} \tag{5.22}$$

□

The sampling formula under neutrality (5.2) and that at the end of a sweep (5.9) lead directly to a likelihood-ratio test of neutrality to detect a recent sweep, which compares the null hypothesis of neutrality ($\gamma = \infty$) and the alternative hypothesis ($\gamma < \infty$). To perform the likelihood-ratio test, we need to maximize (5.9) for θ and γ. The likelihood ratio test statistic, Λ, is

$$\Lambda = \frac{L(\hat{\gamma}, \hat{\theta})}{L_0}, \tag{5.23}$$

where the denominator is (5.2) and the numerator is (5.9) with replacing γ and θ by their maximum likelihood estimators. Appealing to large sample results, $2 \ln \Lambda \sim \chi^2_2$. By using the sampling formula at the end of a sweep (5.9), it is straightforward to compute powers of the tests considered here. The powers to detect a sweep at the end of the sweep by a sample with $n = 50$ as a function of the mutation rate is shown in Figure 5.3. It is assumed that $\gamma = 0.1$. The number of alleles is set to be the nearest integer to the expectation given by the distribution under the sweep (5.8). It can be seen that powers become higher as the mutation rate become higher, which is consistent with the result obtained by extensive simulations (Zeng et al., 2007). As is shown in Figure 5.3, the likelihood ratio test was generally slightly less powerful than the test whose summary statistics is the frequency of the most common allele.

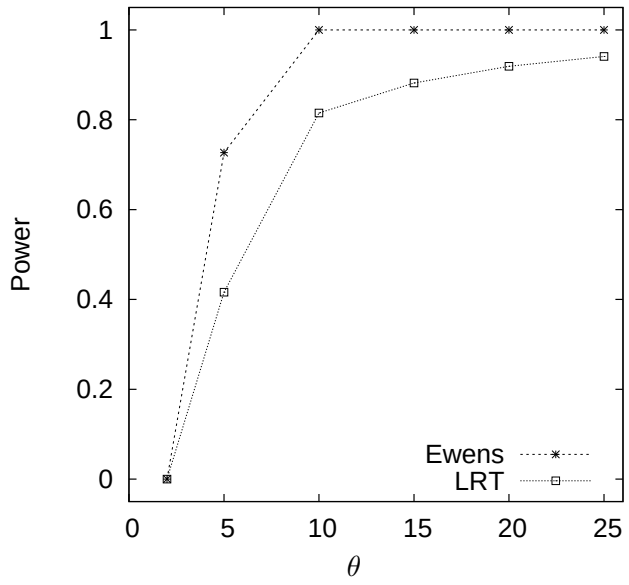

FIGURE 5.3. Powers of the Ewens test and the likelihood ratio test as a function of the mutation rate. $n = 50,\ \gamma = 0.1$.

5.4. Summary

An approximate sampling formula for the infinite allele model at the end of a selective sweep is obtained, in terms of a weighted binomial mixture of the Ewens sampling formula. The probability of mixture is firstly obtained by Maynard Smith and Haigh (1974), and the approximate sampling formula is accurate with probability $1 - O(1/\log \alpha)$. Although the approximation is clued, the formula will give a simple and useful framework for theoretical understanding of allele frequency distribution at the end of a sweep. Barton (1998) pointed out that distribution of "families" of lineages, each sharing a different common ancestor at approximately the same time (at the recombining into the b background), have plenty information. He pointed out that a population size bottleneck could readily distinguished from a sweep if one knew homozygosity and number of families (Figure 8 and 9 in Barton (1998)). In the approximated sampling formula presented here, a family consists of non recombinants alleles and other singleton families in his terminology. However, these singleton could be related to each other and make up some families size of which is larger than one. In this sense, the Watterson homozygosity test could be regarded as a way to detecting sweep by using both of the homozygosity and number of families.

By using the approximate sampling formula for the infinite allele model at the end of a sweep, the new likelihood based test to detect recent selective sweep is presented. The test seems slightly less powerful than the test based on the frequency of the most common

allele when the mutation rate (size of the neutral region) is low, however, the test gives estimates of γ. γ involves two parameters: the selection coefficient and the recombination fraction between the loci. Nevertheless, these two parameters could be easily disentangled if we know genetic map of a region. Because the test is based on a single likelihood, the test could be done as standard likelihood ratio test. It is possible to contract a test based on the composite likelihood. The test should be done by using simulations, but the gain of power is attractive. Finally, assessment of whether the likelihood based test presented here is still powerful and robust when it is applied to haplotype data with intra-haplotype recombination should be addressed.

Bibliography

Athreya, S. R. and Swart, J. M. (2005) Branching-coalescing particle systems. *Prob. Theory Relat. Fields* **131** 376–414.

Barton, N. H. (1999) The effect of hitch-hiking on neutral genealogies. *Genet. Res. Camb.* **72** 123–133.

Brown, T. C. and Jiricny, J. (1987) A specific mismatch repair event protects mammalian cells from loss of 5-methylcytosine. *Cell* **50** 945–950.

Chakraborty, R. and Weiss, K. M. (1988) Admixture as a tool for finding linked genes and detecting that difference from allelic association. *Proc. Natl. Acad. Sci. USA* **85** 9119–9123.

Comtet, L. (1974) *Advanced Combinatrics.* D. Reidel Publishing., Boston.

Depaulis, F. and Veuille, M. (1998) Neutrality tests based on the distribution of haplotyes under an infinite-site model. *Mol. Biol. Evol.* **15** 1788–1790.

Dover, G. (1982) Molecular drive: a cohesive mode of species evolution. *Nature* **299** 111–117.

Durret, L., Mouchiroud, D. and Galtier, C. (1995) Statistical analysis of vertebrate sequences reveals that long genes are scares in GC-rich isochores. *J. Mol. Evol.* **40** 308–317.

Erdélyi, A. (Ed.) (1953) *Higher Transcendental Functions, Vol. I.* McGraw-Hill.

Erdélyi, A. (Ed.) (1954) *Tables of Integral Transforms, Vol. II.* McGraw-Hill.

Etheridge, A., Pfaffelhuber, P., Wakolbinger, A. (2006). An approximate sampling formula under genetic hitchhiking. *Ann. Appl. Prob.* **16** 685–729.

Ethier, S. N. (1979) A limit theorem for two-locus diffusion models in population genetics. *J. Appl. Prob.* **16** 402–408.

Ethier, S. N. and Griffiths, R. C. (1990) On the two-locus sampling distribution. *J. Math. Biol.* **29** 131–159.

Ethier, S. N. and Kurtz, T. G. (1986) *Markov Processes.* Jonh Wiley & Sons.

Ethier, S. N. and Nagylaki, T. (1989) Diffusion approximation of the two-locus Wright-Fisher model. *J. Math. Biol.* **27** 17–28.

Ewens, W. J. (1972) The sampling theory for selectively neural alleles. *Theor. Pop. Biol.* **3** 87–112.

Ewens, W. J. (1973a) Conditional diffusion process in population genetics. *Theor. Pop. Biol.* **4** 21–30.

Ewens, W. J. (1973b) Testing for increased mutation rate for neutral alleles. *Theor. Pop. Biol.* **4** 251–258.

Fay, J. C. and Wu, C.-I. (2000) Hitchhiking under positive darwinian selection. *Genetics* **155** 1405–1413.

Fearnhead, P. (2002) The common ancestor at a nonneutral locus. *J. Appl. Prob.* **39** 38–54.

Feller, W. (1952) The parabolic differentia equations and the associated semi-groups of transformations. *Ann. Math.* **55** 468–519.

Fisher, R. A. (1930) The distribution of gene ratios for rare mutations. *Proc. Roy. Soc. Edinb.* **50** 205–550.

Flammer, C. (1957) *Spheroidal Wave Functions*. Stanford University Press, Stanford.

Fu, Y.-X. and Li, W.-H. (1993) Statistical tests of neutrality of mutations. *Genetics* **133** 693–709.

Galtier, N., Piganeau, G., Mouchiroud, D., Duret, L. (2001) GC-content evolution in mammalian genomes: The biased gene conversion hypothesis. *Genetics* **159** 907–911.

Galtier, N. (2003) Gene conversion drives GC content evolution in mammalian histones. *Trends Genet.* **19** 65–68.

Gantmacher, F. R. (1959) *Matrix Theory, Vol. II*. Chelsea, New York.

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

Griffiths, R. C. (1979) Exact sampling distributions from the infinite neutral allele model. *Adv. Appl. Prob.* **11** 326–354.

Griffiths, R. C. (1980) Line of descent in the diffusion approximation of neutral Wright-Fisher models. *Theor. Pop. Biol.* **17** 37–50.

Griffiths, R. C. (1981) Neutral two-locus multiple allele model with recombination. *Theor. Popul. Biol.* **19** 169–186.

Griffiths, R. C. (1991) The two-locus ancestral graph, in "Selected Proceedings of the Symposium on Applied Probability, Sheffield, 1989" (I. V. Basawa and R. L. Taylor, Eds.), pp. 100–117, IMS Lecture Notes–Monograph Series, Vol. 18, Institute of Mathematical Statistics.

Hill, W. G. and Robertson, A. (1966) The effect of linkage on limits to artificial selection. *Genet. Res. Camb.* **8** 269–294.

Hill, W. G. and Robertson, A. (1968) Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* **38** 226–231.

Hillis, D., Moritz, C., Porter, C. and Baker, R. (1991) Evidence for biased gene conversion in concerted evolution of ribosomal dna. *Science* **251** 308–310.

Hinch, E.J. (1991) *Perturbation Methods.* Cambridge University Press.

Hudson, R. R. (1983a) Property of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* **23** 183–201.

Hudson, R. R. (1983b) Testing the constant-rate neutral allele model with protein sequence data. *Evolution* **37** 213–217.

Hudson, R. R. (1985) The sampling distribution of linkage disequilibrium under an infite allele model without selection. *Genetics* **109** 611–631.

Innan, H. (2002) A method for estimating the mutation, gene conversion and recombination parameters in small multigene families. *Genetics* **161** 865–872.

Innan, H. (2003a) The coalescent and innite-site model of a small multigene family. *Genetics* **163** 803–810.

Innan, H. (2003b) A two-locus gene conversion model with selection and its application to the human RHCE and RHD genes. *Proc. Natl. Acad. Sci. USA* **100** 8793–8798.

Innan, H., Zhang, K., Majoram, P., Tavaré, S., Rosenberg, N. A. (2005) Statistical tests of the coalescent model based on the haplotype frequency distribution and the number of segregating sites. *Genetics* **169** 1763–1777.

Jukes, T. H. and Cantor, D. R. (1969) Evolution of protein molecules, in "Mammalian Protein Metabolism", (H. N. Munro Ed.), pp. 21–132. Academic Press, New York.

Kaplan, N. L. and Weir, B. S. (1992) Expected behavior of conditional linkage disequilibrium. *Am. J. Hum. Genet.* **51** 333–343.

Karlin, S. and McGregor, J. (1968) Rates and probabilities of fixation for two locus random mating finite population without selection. *Genetics* **58** 141–159.

Karlin, S. and Taylor, H. M. (1975). *A First Course in Stochastic Processes, 2nd. ed.* Academic Press, San Diego.

Kim, Y. and Stephan, W. (2002) Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* **160** 765–777.

Kimmel, M., Chakraborty, R., King, J. P., Bamshad, M. Watkins, W. S. and Jorde, L. B. (1998) Signatures of population expansion in microsatellites repeat data. *Genetics* **148**

1921–1930.

Kimura, M. (1955a) Solution of a process of random genetic drift with a continuous model. *Proc. Natl. Acad. Sci. USA* **41** 144–150.

Kimura, M. (1955b) Random genetic drift in multi-allelic locus. *Evolution* **9** 419–435.

Kimura, M. (1955c) Stochastic process and distribution of gene frequencies under natural selection. *Cold Spring Harbor Symposia on Quantitative Biology* **20** 33–53.

Kimura, M. (1957) Some problems of stochastic processes in genetics. *Ann. Math. Stat.* **28** 882–901.

Kimura, M. (1968) Evolutionary rate at the molecular level. *Nature* **217** 624–626.

Kimura, M. and King, J.L. (1979) Fixation of a deleterious allele at one of two "duplicate" loci by mutation pressure and random drift. *Proc. Natul. Acad. Sci. USA* **76** 2858–2861.

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

Kimura, M. and Ohta, T. (1971) *Theoretical Aspects of Population Genetics.* Princeton University Press, Princeton.

Kingman, J. F. C. (1982a) On the genealogy of large populations. *J. Appl. Probab.* **19** 27–43.

Kingman, J. F. C. (1982b) The coalescent. *Stochastic Process. Appl.* **13** 235–248.

Krone, S. M. and Neuhauser, C. (1997) Ancestral process with selection. *Theor. Pop. Biol.* **51** 210–237.

Kruglyak, L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* **22** 139–144.

Laan, M. and Pääbo, S. (1997) Demographic history and linkage disequilibrium in human populations. *Nat. Genet.* **17** 371–373.

Lander, E.S. and Botstein, D. (1986) Mapping complex genetic traits in humans: new methods using a complete RFLP linkage map. *Cold spring harbor symp. quant. biol.* **51** 49–62.

Li, W-H. (1977) Distribution of nucleotide difference between two randomly chosen cistrons in a finite population. *Genetics* **85** 331–337.

Litter, R. A. (1973) Linkage disequilibrium in two-locus, finite, random mating models without selection or mutation. *Theor. Popul. Biol.* **4** 259–275.

Mano, S. (2005) Random genetic drift and gamete frequency. *Genetics* **171** 2043–2050.

Mano, S. (2006) Ewens sampling formula under selective sweep. Recent Advances in Statistical Genetics and Bioinformatics, Issac Newton Institute for Mathematical Science,

Cambridge. poster

Mano, S. (2007) Evolution of linkage disequilibrium of the founders in exponentially growing populations. *Theor. Popul. Biol.* **71** 95–108.

Mano, S. Ancestral process and diffusion model with selection. `arXiv:0804.2696`

Mano, S. and Innan, H. The evolutionary rate of duplicated genes under concerted evolution. *Genetics* 180, in press.

Maruyama, T. (1971) On the fixation probability of mutant genes in a subdevided population. *Genet. Res. Camb.* **15** 221–225.

Maruyama, T. (1972) The average number and the variance of generations at particular gene frequency in the course of fixation of a mutant gene in a finite population. *Genet. Res. Camb.* **19** 109–113.

Maruyama, T. and Kimura, M. (1974) A note on the speed of gene frequency changes in reverse directions in a finite population. *Evolution* **28** 161–163.

Maruyama, T. and Takahata, N. (1981) Numerical studies of the frequency trajectories in the process of fixation of null genes at duplicated loci. *Heredity* **46** 49–57.

Maruyama, T. (1982) Stochastic integrals and their application to population genetics. pp 151-166. In: *Molecular Evolution, Protein Polymorphism and the Neutral Theory*, Edited by M. Kimura. Springer-Verlag. Berlin.

Maynard Smith, J. and Haigh, J. (1974) The hitch-hiking effect of a favourable gene. *Genet. Res. Camb.* **23** 23–35.

Nagylaki, T. and Petes, T. (1982) Intrachromosomal gene conversion and the maintenance of sequence homogeneity among repeated genes. *Genetics* **100** 315–337.

Nei, M. and Li, W-H. (1980) Non-random association between electromorphs and inversion chromosomes in finite populations. *Genet. Res., Camb.* **35** 65–83.

Nei, M., Maruyama, T. and Chakraborty, R. (1975) The bottleneck effect and genetic variability in populations. *Evolution* **29** 1–10.

Nielsen, R. Williamson, S., Kim, Y., Hubisz, M. J., Clark, A. G. and Bustamante, C. (2006) Genomic scans for selective sweeps using SNP data. *Genome Res.* **15** 1566–1575.

Oleinik, O. A. and Radkevic, E. V. (1973) *Second Order Equations with Nonnegative Characteristic Form*. American Mathematical Society, Provine.

Ohta, T. (1980) *Evolution and Variation of Multigene Families*. Springer-Verlag, Berlin/New York.

Ohta, T. (1983a) On the evolution of multigene families. *Theor. Popul. Biol.* **23** 216–240.

Ohta, T. (1983b) Time until fixation of a mutant belonging to a multigene family. *Genet. Res. Camb.* **41** 47–55.

Ohta, T., and Kimura, M. (1969a) Linkage disequilibrium due to random genetic drift. *Genet. Res. Camb.* **13** 47–55.

Ohta, T., and Kimura, M. (1969b) Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. *Genetics* **63** 229–238.

Øksendal, B. (2000) *Stochastic Differential Equations, 5th Ed.* Springer-Verlag. Berlin, Heidelberg.

Pritchard, J. K. and Przeworski, M. (2001) Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69** 1–14.

Przeworski, M. and Wall, J. D. (2001) Why is there so little intragenic linkage disequilibrium in humans? *Genet. Res. Camb.* **77** 143–151.

Reich, D. E. and Goldstein, D. B. (1998) Genetic evidence for a Paleolithic human population expansion in Africa. *Proc. Natl. Acad. Sci. USA* **95** 8119–8123.

Rogers, A. R. and Harpending, H. (1992) Population growth makes waves in the distribution of pairwise genetic differences. *Mol. Biol. Evol.* **9** 552–569.

Sabeti et al. (19 co-authors) (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419** 832–837.

Shiga, T. (1981) Diffusion processes in population genetics. *J. Math. Kyoto Univ.* **21** 133–151.

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

Slatkin, M. (1994a) Linkage disequilibrium in growing and stable populations. *Genetics* **137** 331–336.

Slatkin, M. (1994b) An exact test for neutrality based on the Ewens sampling distribution. *Genet. Res.* **64** 71–74.

Stratton, J. A., Morse, P. M. Chu, L. J. and Hunter, R. A. (1941) *Elliptic Cylinder and Spheroidal Wave Functions.* John Wiley and Sons, New York.

Tavaré, S. (1984) Line-of-descent and genealogical processes, and their applications in population genetics. *Theor. Pop. Biol.* **26** 119–164.

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

Tajima, F. (1989a) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123** 585–595.

Tajima, F. (1989b) The effect of change in population size on DNA polymorphism. *Genetics* **123** 597–601.

Terwilliger, J. D., Zöllner, S., Laan, M. and Pööbo, S. (1998) Mapping genes through the use of linkage disequilibrium generated by genetic drift: 'drift mapping' in small populations with no demographic expansion. *Hum. Hered.* **48** 138–154.

Teshima, K. M., and Innan, H. (2004) The effect of gene conversion on the divergence between duplicated genes. *Genetics* **166** 1553–1560.

Thomas, J. (2006) Concerted evolution of two novel protein families in Caenorhabditis species. *Genetics* **172** 2269-2281.

Thomson, R., Pritchard, J. K. Shen, P. Oefner, P.J. and Feldman, M.W. (2000) Recent common ancestry of human Y chromosomes: Evidence from DNA sequence data. *Proc. Natl. Acad. Sci. USA* **97** 7360–7365.

Walsh, J. B. (1985) Interaction of selection and biased gene conversion in a multigene family. *Proc. Natl. Acad. Sci. USA* **82** 153–157.

Watterson, G. A. (1978) The homozygosity test of neutrality. *Genetics* **88** 405–417.

Wolfram Research, Inc. (2004) Mathematica, Version 5.1, Champaign, IL.

Wright, S. (1931) Evolution in Mendelian populations. *Genetics* **16** 97–159.

Wright, S. (1938) Size of population and breeding structure in relation to evolution. *Science* **87** 430–431.

Wright, S. (1945) The differential equation of the distribution of gene frequencies. *Proc. Nat. Acad. Sci. USA* **31** 382–389.

Zeng, K. Mano, S., Shi, S. and Wu, C.-I. (2007) Comparisons of site- and haplotype-frequency methods for detecting positive selection. *Mol. Biol. Evol.* **24** 1562–1574

Zuckerkandl, E. and Pauling, L. (1965) Evolutionary divergence and convergence in proteins, pp. 97–166. in Evolving Genes and Proteins, edited by V. Bryson and H.J. Vogel. Academic Press, New York.