4

Population Genetics of Molecular Evolution

Carlos D. Bustamante

Department of Biological Statistics and Computational Biology, Cornell University, 422 Warren Hall, Ithaca, NY 14850, USA, cdb28@cornell.edu

Summary

The aim of this chapter is to provide an introduction to aspects of population genetics theory that are relevant to current research in molecular evolution. We review the roles of mutation rates, natural selection, ancestral polymorphism, and linkage among sites in molecular evolution. We also discuss why it is possible to detect the workings of natural selection from comparing rates of substitution for different classes of mutations along a branch in the phylogeny. The problem of estimating the distribution of selective effects among newly arising mutations is given considerable treatment, as are neutral, nearly neutral, and selective population genetics theories of molecular evolution. The chapter does not aim to be an exhaustive description of the field but rather a selective guide to the literature and theory of the population genetics of molecular evolution.

4.1 Introduction

Evolution is the outcome of population-level processes that transform genetic variation within species into genetic differences among species in time and space. Two central goals of evolutionary biology are to describe both the branching order of the history of life (phylogeny) and the evolutionary forces (selective and nonselective) that explain why species differ from one another. Since the 1980s there has been an explosion in the number and complexity of probabilistic models for tackling the first problem, with the motivation that to understand evolution at any level one needs to get the history right (or at least integrate over one's uncertainty in the matter) (for a review, see [113]). Current Markov chain models of evolution deal with the complexities of DNA [48, 60, 40, 119]), RNA [69, 92], codon [39, 70], and protein evolution (see [104] for a review), as well as rate variation among sites [120, 26] and diverse complex dependencies such as tertiary structure [85] and CpG mutational effects [94]. Likewise, there has been tremendous growth in using probabilistic models for hypothesis testing and model selection. For example, it is currently possible to exploit rate variation among codons [72, 124] and among lineages and codons [123] to detect amino acid sites that are likely to be involved in adaptive evolution assuming silent sites evolve neutrally and codons evolve independently of one another.

The purpose of this chapter is to introduce population genetics concepts relevant to the study of molecular evolution, with particular emphasis on understanding how natural selection affects rates and patterns of molecular evolution. Some effort is also made to discuss how population genetics models relate to continuous-time discrete-space Markov chain models of molecular evolution. For example, if the transformation of genetic variation is mostly governed by genetic drift acting on evolutionarily neutral mutations that evolve independently of one another, the outcome will be a Poisson process with constant rate that is independent of the species size [81, 88, 51]. A Markov chain model of evolution (perhaps with rate variation among sites) is a quite appropriate model to capture the dynamics of such a system since the exponential distribution of times among transitions corresponds to an underlying Poisson process. If mutations are not neutral but sites evolve independently of one another, the substitution process can remain a Poisson process that differs among lineages depending on population size and the strength of selection. Under such a model, it is possible to use variation in the rates of substitution among sites to infer the distribution of selective effects among new mutations [25, 73, 90]. Alternatively, if mutations are linked and either slightly deleterious or advantageous (e.g., [81, 77, 78, 79, 59]), or if the fitness effects of mutations vary randomly with the environment (e.g., [100, 30, 31]), the observed patterns of molecular evolution can depart greatly from the expectations of a Poisson process with constant rate [31, 32, 33, 34, 17, 18].

We will begin with a brief historical overview of the population genetics of molecular evolution (Subsection 4.1.1). In Section 4.2, we discuss some of the major predictions of neutral and nearly neutral models of molecular evolution. In Section 4.3, we demonstrate how the classical Wright-Fisher models of population genetics give rise to the neutral theory of molecular evolution. Next will follow a discussion on how ancestral polymorphism can cause departures from the expectations of the neutral independence-among-sites model (Section 4.4). We will then discuss natural selection and demonstrate how comparing the rate of substitution of a putatively selected class of mutations to a neutrally evolving class can be used to infer the signature of natural selection from sequence data (Section 4.5). A discussion will follow on the effects of a distribution of selection coefficients among new mutations on rates and patterns of molecular evolution. Lastly, we investigate the effects of linkage and selection on rates of molecular evolution. A definitive and more mathematical treatment of the subject of theoretical population genetics can be found in Warren Ewens' excellent work *Mathematical Population Genetics*, which has just been published in a second edition by Springer [23].

4.1.1 Setting the Stage

To understand the relationship between population genetics and the study of molecular evolution, one must begin at the point in history where the two became intertwined. In their seminal paper, Zukerkandl and Pauling [126] proposed that the preferred characteristic for inferring the evolutionary relationships among organisms ought to be similarity at the level of DNA or protein sequences. Their paper, while deeply philosophical and contentious, was rooted in the observation that the rate of amino acid evolution in hemoglobin-α and cytochrome-c per year was roughly constant for various vertebrate species. If DNA and protein sequences ("informational macromolecules") accrued substitutions at a near constant rate, then the changes along the phylogeny represented a "molecular clock" that could be used for dating species divergence. Since these changes are more plentiful and presumably subject to less scrutiny by natural selection than morphological characters, the authors reasoned that DNA and protein changes provide better markers for inferring evolutionary relationships. Their paper provided a simple stochastic model of molecular evolution whereby each site had equal probability of being substituted and the number of substitutions that occur along a branch was proportional to the length.

The theoretical foundation for this model (and thus for the molecular clock hypothesis and ultimately for modern-day methods) was provided by the "neutral-mutation drift" theory of molecular evolution, which posited that the vast majority of molecular evolution was due to the stochastic fixation of selectively neutral mutations [55, 63, 57, 62]. The theory concerns both variation within and between species and is summed up most elegantly by the title of Kimura and Ohta's seminal paper: "Protein polymorphism as a phase of molecular evolution" [62]. In other words, the neutral theory arises from considering the evolutionary implications of genetic drift operating on neutral variation [55, 62, 58]. As we will see, the theory predicts (among other things) that the rate of molecular evolution ought to be independent of the population size. In many ways, the true concern of the theory is the distribution of selective effects among newly arising mutations since everything else follows from this premise. The neutral theory is predicated upon the notion that almost all mutations are either highly deleterious or evolutionarily neutral. Highly deleterious mutations contribute little to variation within species and nothing to the genetic differences among species. Adaptive mutations are assumed to be very rare and to fix quickly, thus leaving neutral mutations as the only real source of genetic variation within species that can lead to fixed differences among species. It is important to note that the mature theory says little about the proportion of all mutations that are neutral; rather, it states that most mutations that go on to contribute to differences among species and variation within species are neutral. In this sense, even very constrained molecules such as histones can evolve neutrally. Their molecular clock just ticks at a much lower rate than that of unconstrained molecules such as, perhaps, noncoding DNA. Present-day rate-variation models [120, 26] allow this constraint parameter to vary among sites.

While the neutral theory arises as an extension of population genetics theory, it is not the only population genetics theory of molecular evolution (e.g., [81, 100, 79, 30, 59, 82, 35, 89, 90]). In fact, the field of population genetics has had a long-standing debate over the relative contribution of competing evolutionary forces (mutation, migration, genetic drift, and natural selection) to patterning genetic differences among species. Much of this debate has focused on the question of how much genetic variation within species is maintained by natural selection as well as how much of the molecular differences that we observe among species are due to adaptive molecular evolution [64, 61, 31].

One of the most important critiques of the neutral theory has been put forth by John Gillespie in *The Causes of Molecular Evolution* [31]. He used two lines of evidence to argue that most amino acid substitutions are adaptive. The first is specific examples of adaptive molecular evolution in response to environmental stress. The second is a thorough analysis of variation in the index of dispersion (ratio of the variance to the mean) for amino acid substitutions among mammalian and *Drosophila* species. As mentioned above, a major prediction of the neutral model is that the pattern of substitutions along different branches in a phylogeny ought to be Poisson-distributed with constant rate [81]. Gillespie conclusively demonstrated that the index of dispersion is, on average, much greater than 1 for both sets of species (i.e., it is overdispersed) and that the observations cannot easily be accounted for by neutral or nearly neutral models. He concludes that amino acid evolution occurs due to natural selection in "response to environmental factors, either external or internal, that are changing through time/or space." While the specific model Gillespie espoused [30] may not explain the overdispersed molecular clock (see [34, 35, 17, 18]), the data are certainly not consistent with the strict neutral model.

In fact, recent genome-wide analyses suggest quite an important role for both adaptive and weak negative natural selection in patterning molecular evolution in *Drosophila* (e.g., [91, 24, 90, 75, 98, 5, 8, 90, 38, 93, 6, 84]), *Arabidopsis* (e.g., [8, 67, 4, 110, 84]), maize (e.g., [103, 14, 47]), mouse [96], HIV (e.g., [118, 115, 121, 125, 68, 12, 19]), mammalian mitochondrial genomes [73, 112], and humans (e.g., [46, 87, 83, 1, 41, 13, 97, 29, 50, 114]). While many agree selection is important, there is still considerable debate as to the relative contribution of negative versus positive selection in patterning molecular evolution. As we will see in Section 4.6, the key to the debate rests on rates of recombination and the distribution of selective effects among newly arising mutations. In the next section, we will delve into the specifics of neutral and nearly neutral models before turning to the underlying population genetics machinery.

4.2 The Neutral Theory of Molecular Evolution

It is Darwin [20], of course, who posited that evolution occurs as the result of natural selection by which heritable differences that alter the probability of survival and reproduction of organisms are passed on from generation to generation. Sir Ronald Fisher [27, 28] and Sewall Wright [116] provided the first mathematical models of "the Darwinian evolution of Mendelian populations" by treating genetic drift (i.e., fluctuations in allele frequencies at a given locus due to finite population size) as analogous to the diffusion of heat along a metal bar. In these works, Wright and Fisher also provided the first genetic theories of evolution by deriving a formula for the probability that a mutation subject to natural selection would become fixed in the population (a result we will derive in Section 4.3). What they showed is that if a mutation alters the expected number of offspring a haploid individual (chromosome) contributes to the next generation by a small amount s so that those carrying the mutation leave on average $1 + s$ offsprings and those that do not carry the mutation leave 1 offspring on average, then the probability that a new mutation eventually becomes fixed in the population is roughly

$$\Pr(\text{fixation}) \approx \frac{2s}{1 - \mathrm{e}^{-4Ns}}, \tag{4.1}$$

where N is the effective population size of the species, $2N$ is the number of chromosomes in the population, and s is on the order of N^{-1}. If $s > 0$, we say the mutation is *selectively favored* and that there is *positive selection* operating on the mutation since as the magnitude of s increases above 0 so does the probability of fixation (4.1). Likewise, if $s < 0$, we say the mutation is *selectively disfavored* and there is *negative selection* operating on the mutation since as s becomes more negative, the probability of eventual fixation becomes smaller and smaller. In the neutral case ($s \approx 0$), we can see by applying L'Hopital's rule that the probability of eventual fixation is simply the initial frequency of the mutation $p = \frac{1}{2N}$ (the mutation must have occurred in a heterozygous form).

While Fisher and Wright laid out a great deal of the foundation, it is Motoo Kimura who built up much of the population genetics theory of molecular evolution. His neutral theory of molecular evolution [55, 57, 58, 61] arises from a beautifully simple cancellation of terms: if mutations enter the population at some rate μ per locus per generation, some fraction f_0 are neutral, and $1 - f_0$ are completely lethal, then the rate of evolution k_0 would equal the neutral mutation rate:

$$\begin{aligned} k_0 &= E(\#\text{ of neutral mutations entering per generation.}) \qquad (4.2)\\ &\quad \times \Pr(\text{neutral mutation becomes fixed})\\ &= 2Nf_0\mu\frac{1}{2N}\\ &= f_0\mu\,. \qquad (4.3) \end{aligned}$$

Three major predictions or consequences arise from (4.3):

1. Neutral molecular evolution is independent of the population size and depends only on the *per generation* rate of input of neutral mutations.
2. Neutral molecular evolution is linear in time, thus providing a "molecular clock" by which the relative divergence time of different populations can be dated.
3. Since neutral evolution occurs more rapidly in regions of low selective constraint (high f_0) and more slowly in regions of high selective constraint (low f_0), differences in rates of substitution can be used to infer functional constraint [63].

Furthermore, it is often assumed that the number of neutral mutations that fix in some interval of t generations (substitutions) is Poisson-distributed with rate $k_0 t$.

Our goal in Section 4.3 is to understand the population genetics theory behind equation (4.3) and, more importantly, to understand when this simple neutral model holds and when it does not hold. For example, the assertion that the substitution process is a Poisson process only holds if sites evolve independently of one another [51, 108]. This will be true only if there is free recombination among sites or if there is a sufficiently low mutation rate that only 1 or 0 nucleotides vary at a given point in time for a non-recombining region. High mutation rates and linkage among neutral sites can have a pronounced effect, leading to the fixation of "bursts" of mutations that are approximately geometrically distributed [108, 109, 32].

It is important to mention at this point that population genetics models of molecular evolution differ in some regards from discrete-space continuous-time models [48, 40, 60, 119]. For example, the Poisson assertion above ignores the possibility of multiple substitutions at the same site. The reason many population genetics models make such an assumption is that the timescale on which they operate is relatively short compared with the timescale on which phylogenetic reconstruction of distantly related species is usually carried out. Likewise, much of the theory is based on the behavior of single-locus two-allele models, where the goal is to understand the probability of fixation of a new mutation under various scenarios. Such a model is not rooted in the actual A, C, T, and G of DNA but rather on the fact that at a given nucleotide site the probability of having more than two nucleotides segregating is very low. Likewise, if the population size and mutation rates are small, there will be few linked polymorphic sites. Therefore, the independently evolving single-locus model with two alleles is a reasonable place to start in modeling molecular evolution.

4.2.1 Nearly Neutral Models of Molecular Evolution

From the beginning, it was evident that the great power of the neutral theory of molecular evolution lay in its quantitative predictions regarding rates and patterns of molecular evolution. In Kimura's original paper [55], the problems the neutral theory solved were the inordinately high rate of nucleotide evolution inferred from patterns of amino acid evolution [126] as well as the plentiful amounts of amino acid variation within species [43, 65]. According to Kimura's calculations, Darwinian evolution would produce too high a genetic load on the population to account for these patterns; therefore, most of the changes were likely neutral. Likewise, King and Jukes [63] set out to demonstrate that "most evolutionary change in proteins may be due to neutral mutations and genetic drift" by testing some of the predictions of a neutral molecular evolution theory using almost all of the available data in the world on protein, RNA, and DNA sequence variation.[1] One key prediction of the neutral theory was that if proteins were more constrained than genomic DNA, then proteins should evolve at a slower rate. If, on the other hand, proteins were constantly being refined by positive natural selection, then the rate of evolution of proteins would be faster than that of genomic DNA. Using early DNA hybridization experiments coupled with protein sequence information, King and Jukes concluded (rightly) that most proteins evolve at a much slower rate than most regions of genomic DNA. Another key argument they used was a near Poisson fit to the number of substitutions per site across the gene trees of various molecules (globins, cytochrome-c, and immunoglobulin-G light chains).

It was soon pointed out that if the neutral theory of molecular evolution was strictly true, then the rate of amino acid evolution should be proportional to generation time and not chronological time. Kimura and Tomoko Ohta [81] countered with the first "nearly neutral" model of molecular evolution. This model posits that newly arising nonlethal mutations are not strictly neutral ($s \approx 0$) but rather have selection coefficients drawn from a distribution such that the mean selective effect is slightly deleterious and most mutations are in the interval ($-\frac{1}{N} \leq s \leq \frac{1}{N}$).[2] Under such a scheme, the evolutionary fate

[1] King and Jukes had independently proposed a neutral theory of molecular evolution, but their paper was initially rejected by *Science*. In the interim, Kimura's paper appeared, and Kimura's results were added to the revised King and Jukes manuscript [99].

[2] The definition of "nearly neutral" is somewhat of a moving target and context-dependent. In their original paper, Ohta and Kimura [81, p.22] implicitly considered nearly neutral those mutations in the interval ($-\frac{2}{N} \leq s' \leq \frac{2}{N}$), where $s' = 2s$. In Ohta and Kimura's later work [77, 78, 79, 59], the emphasis was on explaining how slightly deleterious mutations could be considered an engine for nonadaptive molecular evolution. Likewise, Gillespie [31] has argued that nearly neutral should only refer to mutations in the interval ($-\frac{1}{N} \leq s' < 0$) since slightly advantageous mutations are helped along by selection. Ohta [80] (not surprisingly) has explicitly reclaimed the "slightly advantageous" as nearly neutral ground by arguing that the fate of slightly advantageous mutations is very much governed by both selection and drift. Unless otherwise noted, we will adopt Ohta's view and consider nearly neutral mutations as those that are in the interval $-2 \leq \gamma \leq 2$, where $\gamma = 2Ns$.

of mutations is mostly governed by genetic drift. One implication of near-neutrality is an inverse relationship between population size N and the rate of molecular evolution at selected sites k_s. Letting f_s be the fraction of mutations that are selected, under the assumption that selected mutations evolve independently of one another, the rate of evolution for a selected mutation k_s is given by

$$\begin{aligned} k_s &= E(\#\text{ of selected mutations entering per generation.}) \qquad (4.4)\\ &\qquad \times \Pr(\text{selected mutation becomes fixed})\\ &= 2N f_s \mu \frac{2s}{1-\mathrm{e}^{-4Ns}}\\ &= f_s \mu \frac{4Ns}{1-\mathrm{e}^{-4Ns}}\ . \qquad (4.5)\end{aligned}$$

We see from (4.5) that for a fixed $s < 0$

$$\lim_{N\to\infty} k_s = 0.$$

The interpretation of this equation is that if mutations are slightly deleterious, a species with a large population size will evolve at a slower rate than a species with a small population size. Ohta and Kimura [81] posited that since population size is roughly inversely proportional to body size and body size is roughly inversely proportional to generation time (i.e., big animals have long times between generations but also live at low densities), these two factors cancel each other out to produce a rate of evolution that is close to linear in chronological time. Kimura [59] later argued that if $-s$ follows a Gamma distribution with mean 1 and shape parameter $\beta = 0.5$, then the rate of evolution will be proportional to $\sqrt{N}$.

A very useful way of studying the consequences of natural selection on rates of molecular evolution is by comparing the relative rate of substitution for selected mutations (4.5) to neutral mutations (4.3)

$$\omega = \frac{k_s}{k_0} = \frac{f_s}{f_0}\frac{2\gamma}{1-\mathrm{e}^{-2\gamma}}$$

letting $\gamma = 2Ns$. We will refer to γ as the scaled selection coefficient, and it will reappear when we derive (4.5) from an approximation to the Wright-Fisher process (Section 4.5). We note that ω can be interpreted as the expected dn/ds ratio assuming silent mutations are neutral, replacement mutations have the same selective effect, and mutations evolve independently of one another. Assuming $f_s = f_0$, if $s = -1 \times 10^{-4}$ and the population size is small ($N = 1000$), the rate of evolution at selected sites is $\omega = 0.81$, the rate of evolution at neutral sites, which we might refer to as a modest reduction. On the other hand, if s does not change and the population size is large ($N = 10,000$), then $\omega = 0.074$ and we would observe a large reduction in the substitution rate. In Figure 4.1, we plot the rate of substitution for selected mutations as compared with neutrality assuming $f_s = f_0$.

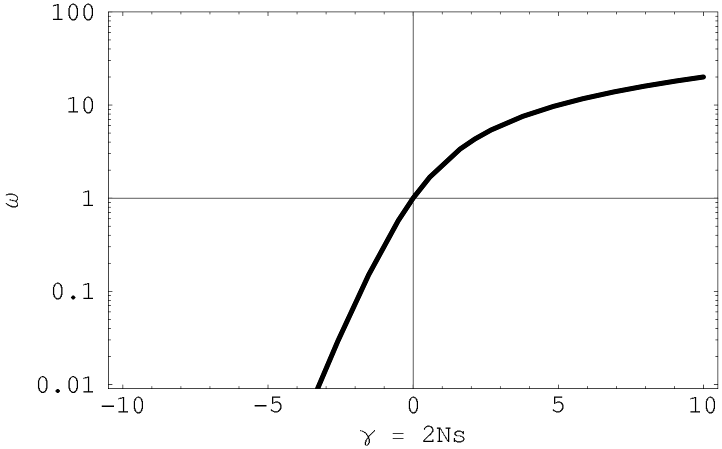


Fig. 4.1. Effect of natural selection on rates of molecular evolution. The x-axis is the scaled selection coefficient for new mutations, and the y-axis is the relative rate of substitution as compared with neutrality. Note that the y-axis is on a log-scale.

4.3 Wright-Fisher Model

4.3.1 No Mutation, Migration, or Selection

Consider a diploid population of constant size N (i.e., a population of $2N$ chromosomes) with discrete nonoverlapping generation [116, 28]. The population in the next generation is produced by randomly pairing gametes from an infinitely large pool of gametes produced by the current population. Focus on a neutrally evolving locus A with two alleles A_1 and A_2, and assume that there is no mutation between A_1 and A_2. Let $X(t)$ be the number of chromosomes in the population that carry the A_1 allele at generation t. The collection of random variables $\{X(t)\}$ for $t = 0, 1, \ldots$ is a discrete-time discrete-space Markov chain with state space $\{0, 1, \ldots, 2N\}$. The transition probability P_{ij} for going from state i to state j comes from binomial sampling:

$$P_{ij} \equiv \Pr(X(t+1) = j \mid X(t) = i) = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2N-j} . \quad (4.6)$$

This model is known as the Wright-Fisher model of population genetics, and the stochastic sampling of gametes from generation to generation is known as *genetic drift*. It is easy to verify that $X(t) = 0$ and $X(t) = 2N$ are *absorbing* states (i.e., $P_{0\,0} = P_{2N\,2N} = 1$), corresponding to loss ($X(t) = 0$) or fixation ($X(t) = 2N$) of the A_1 allele. It is also relatively easy to show that all other states $(1, 2, \ldots, 2N-1)$ are transient. This conforms with our biological intuition that if a population has 0 copies of allele A_1 in generation t_0, $\Pr(X(t) = 0) = 1$ for all $t > t_0$.

An implication of the Wright-Fisher model is that *each segregating neutral mutation in a population is eventually fixed or lost.* The stochastic fixation of neutral mutations (along with the fixation of selected mutations) thus underpins molecular evolution. It is then of immediate interest to find the probability that a mutation initially at frequency $p = \frac{X(0)}{2N}$ is eventually fixed in the population. The expected gene frequency in generation $t+1$ given the gene frequency in generation t comes directly from the binomial model for gametic sampling:

$$E\left(\frac{X(t+1)}{2N} \mid X(t)\right) = \frac{\sum_{j=0}^{2N} jP_{ij}}{2N} = \frac{X(t)}{2N} \; .$$

Similarly, the variance in gene frequency is

$$V\left(\frac{X(t+1)}{2N} \mid X(t)\right) = \frac{X(t)(1-X(t))}{2N} \; .$$

The first result implies that for the Wright-Fisher model without mutation, the expected change in allele frequency from generation to generation is zero (i.e., the $X(t)$ process is a Martingale). We can thus think of the change in gene frequency as a random walk without bias. As a result, we might intuit from symmetry alone that the probability of eventually fixing the A_1 allele should equal the initial frequency of the A_1 allele in the population (i.e., p).

A more rigorous approach is to set up a set of linear recurrence equations that the Wright-Fisher process must satisfy [74, p. 15]. Let p_j be the probability that a population that starts with j copies of the A_1 allele ($X(0) = j$) eventually fixes the A_1 allele (i.e., the probability that the process reaches $2N$ before it reaches 0). Clearly, $p_0 = 0$ and $p_{2N} = 1$. By exploiting the Markov property of the system, we can write down the following set of equations:

$$p_i = \sum_{j=0}^{2N} p_j P_{ij}, \quad \text{for } i = 1, \ldots, 2N-1 \; . \tag{4.7}$$

The reason our model must satisfy these equations is that once the process enters state j, it "forgets" that it had previously been in state i and the process is restarted. The probability of reaching state $2N$ before state 0 is p_j, and by weighing the p_j's by the probability of transitioning from state i into state j, we obtain a set of $2N-1$ equations (4.7) for $2N-1$ unknowns $(p_1, p_2, \ldots, p_{2N-1})$. By substituting (4.6) into (4.7), we verify that $p_j = \frac{j}{2N}$ is the non-negative solution to the system of equations. Therefore, the probability of eventual fixation of a neutral mutation is

$$p_1 = \frac{1}{2N} . \tag{4.8}$$

4.3.2 Rate of Fixation of Neutral Mutations

Now consider a process whereby in each generation a Poisson number of mutations occurs at a rate $\frac{\theta}{2} = 2Nf_0\mu$, where $f_0\mu$ is the generation neutral mutation rate per locus. It is assumed that each mutation occurs at a previously invariant DNA site [58, 107]. We will now consider the rates and patterns of neutral molecular evolution under two assumptions: (a) complete independence among sites [58, 21, 22, 89] and (b) complete linkage among sites [107].

Independence among sites

Following [21, 89], model the mutation process as starting a Poisson number of new Wright-Fisher processes each generation. Let $X_j(t)$ be the state of the process (frequency) at site j at time t, where t is measured as the time since the mutation at site j originated in the population (i.e., $X_j(0) = \frac{1}{2N}$ for all j). It is assumed that mutations $\{i = 1, 2, \ldots\}$ evolve independently of one another so that X_j processes are i.i.d. Considering some absolute interval of time $(0, T]$, let M_i for $i = 1, 2, \ldots, T$ be the number of mutations that enter the population in generation i that are destined to be fixed. The time of entry of mutations that eventually fix in the population is known as the *origination* process [33, 88, 51]. Since each mutation has probability $p_1 = \frac{1}{2N}$ of eventually fixing in the population and the trajectories $X_1, X_2, \ldots$ are independent of each other, M_i for $i = 1, 2, \ldots, T$ are i.i.d. filtered Poisson random variables with rate

$$\mathbb{E}(M_i) = \frac{\theta}{2} p_1 = 2N\mu f_0 \frac{1}{2N} = \mu f_0 \ .$$

Furthermore, the total number of mutations $K = \sum_{i=1}^{T} M_i$ that enter the population during $(0, T]$ and eventually fix is also a Poisson random variable with rate $\mathbb{E}(K) = \mu f_0 T$ by the additivity property of independent Poisson random variables.

It is important to note that K is not the *actual* number of mutations that fix during the given interval of T generations (known as the *fixation process* [33]) but rather the number of mutations that enter during this interval and *eventually* become fixed. In the case of independently evolving sites, the origination process and the fixation process will have the same distribution as long as the time intervals are exchangeable. An example of when the time intervals would not be exchangeable is a difference in mutation rates for different time intervals.

Complete linkage among sites

Birky and Walsh [7] showed that the expected substitution rate for neutral mutations is not affected by linkage to neutral, deleterious, or advantageous mutations. Here we follow Cutler's discussion of the problem [16] closely to show that the distribution of the number of mutations that ultimately fix in the population remains a filtered Poisson process with rate μf_0 [81]. This was originally shown using reversibility arguments by Sawyer [88] and Kelly [51, p. 158].

Assume that mutations enter at a Poisson process rate $\frac{\theta}{2} = 2Nf_0\mu$, and write $X_j(t)$ for $j = 1, 2, \ldots$ to denote the frequency of the j process at time t since the origination of mutation j. Assume complete linkage among sites and write $f_j(x \mid t)dt$ to denote the $\Pr(X_j(t) = x)$. Let us introduce an indicator variable that tracks whether a given mutation becomes fixed in the population:

$$I_j = \begin{cases} 1 & \text{if mutation } j \text{ fixes in the population} \\ 0 & \text{otherwise.} \end{cases}$$

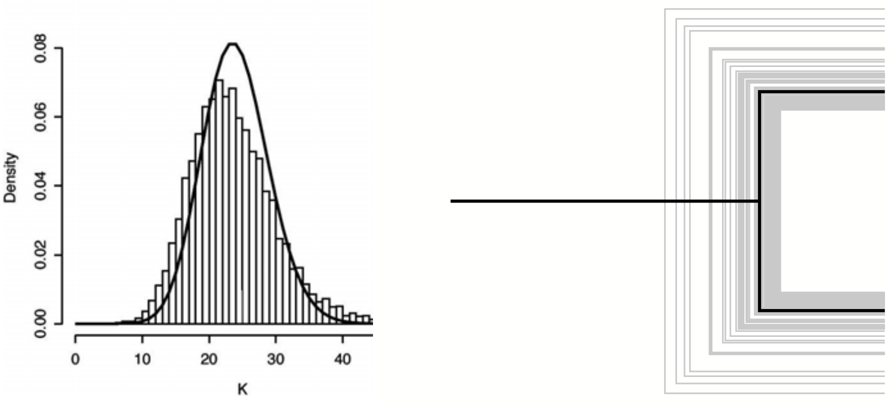
Since the number of neutral mutations on a chromosome does not alter the probability of fixation, $\mathbb{E}(I_j) = p_1$ for all j. Likewise, since the expected change in frequency from generation to generation is 0, the expected frequency of the j process is

$$\mathbb{E}(X_j(t)) = \int_0^1 x f_j(x \mid t)dx = \mathbb{E}(X_j(0)) = p_1 .$$

Now consider two mutations, which we arbitrarily label $j = 1$ and $j = 2$, and assume mutation 1 is older than mutation 2. Consider the probability that both mutations become fixed ($\mathbb{E}(I_1 I_2)$). For this to happen, mutation 2 must occur on a background that contains mutation 1. The probability of this occurring is the frequency of the first mutation at the time the second mutation originates, $X_1(t)$. The marginal probability that mutation 2 fixes is simply its initial frequency $X_2(0) = p_1$. Therefore, the probability that both mutation 1 and mutation 2 fix in the population is given by

$$\begin{aligned} \mathbb{E}(I_1 I_2) &= \Pr(\text{mutation 2 fixes}) \Pr(\text{mutation 1 fixes} \mid \text{mutation 2 fixes}) \\ &= \Pr(\text{mutation 2 fixes}) \cdot \\ &\qquad \Pr(\text{mutation 2 arose on a chromosone containing mutation 1}) \\ &= p_1 \int_0^1 x f_1(x \mid t)dx \\ &= p_1^2 \, . \end{aligned}$$

Since the probability that both mutations fix is shown to be the product of the probability that each mutation fixes alone, the random variables $X_1(t)$ and $X_2(t)$ must be independent. This implies that linkage among neutral mutations does not affect the neutral rate of evolution. Likewise, since X_1 and X_2 are independent, the origination process remains a filtered Poisson process. The fixation process, on the other hand, does not remain a Poisson process in the presence of linkage. Informally, one can reason that the time intervals are no longer exchangeable. As has been discussed by Gillespie [31, 33] and



(a) (b)

Fig. 4.2. Population dynamics can influence of molecular evolution. Two populations are split, evolved for $t = 10N$ generations, and a random chromosome from each is compared. (a) Distribution of the number of differences between a pair of random sequence from two populations that separated $10N$ generations ago and accrue mutations at rate $\mu = \frac{1}{N}$. The solid line is the expected distribution from a Poisson model. (b) Variation in branch length for the process due to random coalescence in ancestral population for $t = 10N$. The black line is the expected branch length (measured along the horizontal axis), and the grey lines are 100 replicates of the process.

Watterson [108, 109], the fixation process for the neutral infinite-sites model is a "burst" process whereby a geometric number of mutations fix when a chromosome reaches frequency 1 in the population. The effect of correlation in the substitution process is to reduce the efficiency of statistical methods for phylogenetic reconstruction [45].

4.4 Ancestral Polymorphism and Neutral Molecular Evolution

The analysis in Section 4.3 is predicated upon being able to follow the history of the entire population. The purpose of this Section is to derive the mean and variance of the sampling distribution for the number of nucleotide differences K between a sample of two DNA sequences drawn from a pair of populations that diverged t generations in the past. The full distribution for a sample of size $n = 2$ can be found in [102].

Measuring time into the past so that 0 is the present day, let

$$K = K_1 + K_2 + K_A,$$

where K_1 and K_2 are the number of mutations that accumulate on the first and second sequences since time t and K_A is the number of fixed differences due to ancestral polymorphism. Assuming a molecular clock, K_1 and K_2 are Poisson with rate $f_0\mu t$. Without loss of generality, assume $f_0 = 1$. It will be shown that K_A is a geometrically distributed random variable so that the sampling distribution of K is not Poisson (see Figure 4.2). We will also see that the degree to which K will differ from a Poisson random variable with the same mean will depend on the parameters t and N_A, where N_A is the ancestral population size.

We will begin by considering the distribution of the number of differences for a sample of two chromosomes drawn from a panmictic population. This is equivalent to deriving the distribution of heterozygosity under an infinite-sites model and is a well-studied problem in population genetics (e.g., [56, 58, 107]). We will use the machinery of coalescent theory [44] to address the issue.

4.4.1 Average Pairwise Distance

Consider a sample of size $n = 2$ chromosomes drawn from a randomly mating population of size $2N$ chromosomes. Let S_2 be the number of nucleotide differences between two sequences at our locus of interest.

The probability that a random pair of chromosomes find a common ancestor in the previous generation is $\frac{1}{2N}$. Therefore, the distribution of the number of generations M until the two chromosomes find a common ancestor is a "first success" distribution with mean $2N$:

$$\Pr(M = m) = \left(1 - \frac{1}{2N}\right)^{m-1} \left(\frac{1}{2N}\right) . \tag{4.9}$$

If N is large, (4.9) can be approximated using an exponential distribution. Measuring time in units of $2N$ generations, the random variable $T_2 = \frac{M}{2N}$ follows the exponential distribution with rate 1,

$$\Pr(M \leq 2Nx) = Pr(T_2 \leq x) \approx 1 - \mathrm{e}^{-x}.$$

The random variable T_2 is known as the *coalescent* time for a sample of size $n = 2$ and describes the waiting time until two random chromosomes from a population coalesce (or merge) in a common ancestor. As one follows the two sequences back in time until the coalescent event, each accrues mutations independently at a rate $\frac{\theta}{2} = 2N\mu$ per unit of time assuming a Poisson model of mutation. This assumption implies that the waiting time until a mutation (T_M) occurs along either chromosome is exponential with rate θ. By the usual result for competing exponentials

$$\Pr(T_M < T_2) = \frac{\theta}{\theta + 1}.$$

Likewise, because of the memoryless property of the exponential distribution, once a mutation event occurs along either chromosome, the coalescent process is restarted. Therefore, the distribution of the number of mutations before a coalescent event for $n = 2$ is geometric:

$$\Pr(S_2 = k) = \left(\frac{\theta}{\theta+1}\right)^k \frac{1}{\theta+1}. \tag{4.10}$$

The expected value and variance of S_2 are easily shown to be

$$\mathbb{E}(S_2) = \theta, \quad \mathbb{V}(S_2) = \theta^2 + \theta. \tag{4.11}$$

Equations (4.10) and (4.11) were first derived by Watterson [107] when he found the distribution of the number of segregating sites S_i in a sample of size i. Li [66] also derived these results while finding the transient distribution of S_2. For our problem, $K_A = S_2$ with N replaced by N_A.

Recall that K is the sum of two independent Poisson random variables, each with mean μt, and a geometric random variable with mean $\theta_A = 4N_A\mu$, where N_A is the size of the ancestral population. This implies that

$$E(K) = 2\mu(t + 2N_A), \quad V(K) = 2\mu(t + 2N_A + 8N_A^2\mu)\ . \tag{4.12}$$

The index of dispersion (the ratio of the variance to the mean) is one way to assess the concordance between K and a Poisson random variable with the same mean [81, 31]. For K it is easy to show that

$$R(K) = 1 + \frac{8N_A^2\mu}{t + 2N_A} = 1 + \frac{\theta_A}{1+\tau}\ ,$$

where $\tau = t/2N_A$. Figure 4.2 illustrates that ancestral polymorphism can lead to deviations from the Poisson expectations. In this figure, we have simulated 10,000 comparisons of $n = 2$ sequences drawn from a pair of populations that diverged $t = 10N_A$ generations ($\tau = 5$) in the past. Mutations occur in each daughter population as a Poisson process with rate $\mu = \frac{1}{N_A}$ per chromosome per generation ($\theta_A = 4$). Note that the distribution of K has a much larger variance than expected from the Poisson prediction ($E(K) = 24$) with $R(K) = 1.666$.

4.4.2 Lineage Sorting

Ancestral polymorphism can also lead to the phenomenon of "lineage sorting", where the genealogical tree for a sample of DNA sequences has a different branching order than the tree relating the history of population-splitting events. That is, if we have a sample of three sequences from three species $\{A, B, C\}$ and the tree relating our three populations is $((A, B), C)$, there is some probability of recovering *discordant* gene trees that are of the form $(A, (B, C))$ and $((A, C), B)$. (For an excellent discussion of the problem from a population genetics perspective, see [86]). The probability of recovering discordant trees in the three-taxon case is relatively easy to calculate using coalescent theory.

Assume that the population size N of three species is the same and has been constant for the history of $\{A, B, C\}$. Let t_1 be the time in the past in units of $2N$ generations when populations A and B split and let t_2 be the time in the past when the ancestral populations of A and B split from C. Write T_{AB} to denote the coalescent time of the sequence from species A and from species B and define T_{AC} and T_{BC} analogously. The probability that a gene tree will be concordant is the probability that A and B coalesce with each other before either coalesces with C. That is, the probability of concordance is given by $\Pr(\min(T_{AB}, T_{AC}, T_{BC})) = T_{AB}$.

The first coalescent event in the history of $\{A, B, C\}$ cannot occur before t_1. Between times t_1 and t_2, only coalescent events between A and B are allowed, and after t_2 all three lineages are equally likely to coalesce with one another. Letting $t = t_2 - t_1$, we can write

$$\begin{aligned} T_{AB} &= t_1 + X_1 \ , \\ T_{BC} &= t_1 + t + X_2 \ , \\ T_{AC} &= t_1 + t + X_3 \ , \end{aligned} \tag{4.13}$$

where X_1, X_2, and X_3 are i.i.d. exponentially distributed random variables with rate 1. The justification for (4.13) comes from the results derived above that for large N the coalescent time for a sample of two sequences is exponential with rate 1. Recalling that the minimum of k independent exponential random variables is exponentially distributed with the sum of the k rates, we can also write

$$\min(T_{BC}, T_{AC}) = t_1 + t + Y \ ,$$

where Y is an exponential random variable with rate 2 that is independent of X_1. Therefore,

$$\begin{aligned} \Pr(\text{concordance}) &= \Pr(\min(t + Y, X_1) = X_1) \\ &= \Pr(\min(t + Y, X_1) = X_1 \mid X_1 \le t) \times \Pr(X_1 \le t) + \\ &\qquad \Pr(\min(X_1, Y) = X_1 \mid X_1 > t) \times \Pr(X_1 > t) \\ &= 1 \times (1 - \mathrm{e}^{-t}) + \frac{1}{3} \times \mathrm{e}^{-t} \\ &= 1 - \frac{2}{3}\mathrm{e}^{-t} \ . \end{aligned}$$

This simple example illustrates that to understand molecular evolutionary patterns on relatively short timescales, one must model the population genetics dynamics.

The question of estimating ancestral population genetics parameters has a rich history. Equations (4.12) were first derived by Takahata and Nei [101]. The full distribution of K in the case of one sequence from each of a pair as well as each of a triplet of species is given in Takahata, Satta, and Klein [102, eqs. (3), (6)]. As they discuss, these probabilities can be used for maximum likelihood estimates of the species divergence time and ancestral population size from multilocus data. Likewise, Yang [122] and Wall [106] have developed methods that incorporate rate variation among loci as well as recombination. The effects of population growth and differences in population size on levels of variation within and between a pair of species are taken up by Wakeley and Hey [105]. Likewise, a Bayesian method for distinguishing migration from isolation using within- and between-species sequence data is presented by Nielsen and Wakeley [71].

4.5 Natural Selection

The Wright-Fisher machinery can be adapted for modeling other evolutionary forces by specifying the joint effects of all forces on the change in gene frequency per generation. This is usually done in a two-step process. First an infinite gamete pool is assumed such that the frequency of the A_2 allele changes in the gamete pool deterministically due to mutation, selection, and other factors from some value $p = \frac{i}{2N}$ to p'. The effect of genetic drift is modeled using an equation analogous to (4.6), where p' depends on i and the evolutionary forces being considered:

$$P_{ij} \equiv \Pr(X(t+1) = j \mid X(t) = i) = \binom{2N}{j} (p')^j \, (1-p')^{2N-j} \ . \tag{4.14}$$

In modeling natural selection, one needs to specify the fitness of the three relevant genotypes. Let the expected relative contribution of the A_1A_1, A_1A_2, and A_2A_2 genotypes to the next generation be 1, $1 + 2sh$, and $1 + 2s$. (Note that h is known as the dominance parameter and summarizes the effect of selection on the heterozygote fitness.) The effect of natural selection is to bias the chance of picking an allele A_2 at random from the next generation. The expected proportion of offspring left by each of the three genotypes is

$$A_1A_1 : \frac{(1-p)^2}{\overline{w}}, \quad A_1A_2 : \frac{2p(1-p)(1+2sh)}{\overline{w}}, \quad A_2A_2 : \frac{(1+2s)p^2}{\overline{w}},$$

where $\overline{w} = (1-p)^2 + 2(1+2sh)p(1-p) + p^2(1+2s)$.

Therefore, the frequency of the A_2 allele after one round of natural selection is

$$p_{t+1} = \frac{p_t^2(1+2s) + (1+2sh)p_t(1-p_t)}{\overline{w}} \ .$$

As we will see below, the number of selected mutations that fix in the history of a population under the assumption of recurrent mutation and selection is also Poisson and depends on the parameter $\gamma = 2Ns$ and h.

4.5.1 Diffusion Approximation

To study the Wright-Fisher model with selection (and other complicated population genetics models), it is often more convenient to work with a continuous-time continuous-space approximation to a discrete process. The natural state space is the frequency of a mutation ($0 \leq x = \frac{X(\cdot)}{2N} \leq 1$), and the natural time scaling is in units of $2N$ generations. Fisher [27] first noted that the action of genetic drift on a locus could be modeled using the same differential equations used to model the diffusion of heat. The classical problem of finding the stationary distribution of allele frequencies visited by a mutation under a variety of selective, mutation, and demographic models was taken up by Fisher in *The Genetical Theory of Natural Selection* [28] as well as by Sewall Wright [116, 117]. The time-dependent solution of what was later recognized as the Fokker-Planck or Kolmogorov forward equation was given in [52]. A definitive treatment of the subject is given in Kimura's classic paper [54]. We will now proceed to derive the stationary distribution, omitting many technical details that can be found by the interested reader in [54, 49, 23].

As discussed in Karlin and Taylor [49, p. 180], as $N \to \infty$, the Wright-Fisher process has a limiting diffusion that depends on the mean $M_{\delta x}$ and variance $V_{\delta x}$ of the change of gene frequency per generation. $M_{\delta x}$ will usually depend on the specifics of the model that produces the change in the gamete pool (mutation, migration, selection, etc.), while $V_{\delta x}$ is almost always given by the effects of binomial sampling. It is important to note that neither $M_{\delta x}$ nor $V_{\delta x}$ depend on time.

Write $\phi(x \mid p, t)dx$ to represent the conditional probability that a mutation at frequency p goes to frequency x in time t. In this equation, p is fixed and x is a random variable. When $dx = \frac{1}{2N}$ is substituted, $f(x \mid p, t) = \phi(x \mid p, t)\frac{1}{2N}$ gives the approximate frequency of mutations in the interval $x + dx$ for $0 < x < 1$ [54]. As discussed in [54], $\phi(x \mid p, t)$ is the solution to the Kolmogorov forward equation

$$\frac{\partial \phi(x \mid p,t)}{\partial t} = \frac{1}{2}\frac{\partial^2}{\partial x^2}\{V_{\delta x}\phi(x \mid p,t)\} - \frac{\partial}{\partial x}\{M_{\delta x}\phi(x \mid p,t)\} \ . \tag{4.15}$$

A very useful consequence of (4.15) is that we can solve for the *stationary* or time-independent solution (if it exists) of $\phi(x \mid p)$ by setting $\frac{\partial \phi(x|p,t)}{\partial t} = 0$,

$$\phi(x) = \frac{C}{V_{\delta x}} \exp\left(-2\int \frac{M_{\delta x}}{V_{\delta x}}\right) \ , \tag{4.16}$$

where C is a constant chosen so that $\int \phi(x)dx = 1$. The time-independent solution of (4.15) was first found by Sewall Wright [117].

Example 4.1: Reversible mutation neutral model

Consider a neutral model with reversible mutation so that $A_1 \to A_2$ at rate μ and $A_2 \to A_1$ at rate ν per generation. Let x_t represent the frequency of the A_1 allele at time t,

$$x_{t+1} = (1 - x_t)\nu - x_t\mu \ ,$$

implying that $M_{\delta x} = (1 - x)\nu - x(1 + \mu)$. The variance of the change in gene frequency is

$$V_{\delta x} = \frac{x(1-x)}{2N} \ .$$

Plugging M_{δ_x} and V_{δ_x} into (4.16), it is relatively straightforward to show that

$$\phi(x) = Cx^{4N\nu-1}(1-x)^{4N\mu-1} \ .$$

Recognizing that this is the density of a Beta distribution with parameters $4N\nu$ and $4N\mu$, the necessary constant is $C = \frac{\Gamma(4N\nu+4N\mu)}{\Gamma(4N\nu)\Gamma(4N\mu)}$.

4.5.2 Probability of Fixation

One of the most useful applications of the diffusion approximation is to calculate the probability of fixation of a mutation given its frequency in the population. To do so, we will follow [53] and use the Kolmogorov backwards equation to solve for $\phi(x \mid p, t)$. In this equation, we write the differential equation with respect to p varying, and the model is equivalent to running the process backwards in time (i.e., reversing the diffusion from x to p). The Kolmogorov backwards equation is

$$\frac{\partial \phi(x \mid p,t)}{\partial t} = \frac{V_{\delta p}}{2}\frac{\partial^2 \phi(p \mid x,t)}{\partial p^2} + M_{\delta p}\frac{\partial \phi(x \mid p,t)}{\partial p} \ . \tag{4.17}$$

If we substitute in $x = 1$, the solution to equation (4.17) gives us the probability of a mutation reaching fixation by time t given an initial frequency p. We will follow Kimura [54] and refer to this probability as $u(p,t)$. The boundary conditions for solving (4.17) are $u(0,t) = 0$ (i.e., probability of reaching 1 before 0 is 0 if $p = 0$) and $u(1,t) = 1$.

Again, following [54], by letting t tend towards infinity, we can find the probability of ultimate fixation:

$$u(p) = \lim_{t\to\infty} u(p,t) \ .$$

For the probability of ultimate fixation, $u(p)$, the left-hand side of (4.17) is 0, and thus the solution satisfies

$$0 = \frac{V_{\delta p}}{2}\frac{d^2 u(p)}{dp^2} + M_{\delta p}\frac{du(p)}{dp} \ .$$

Kimura [53] showed that the solution to this equation is

$$u(p) = \frac{\int_0^p G(x)dx}{\int_0^1 G(x)dx} \ ,$$

where

$$G(x) = \exp\left(-2\int \frac{M_{\delta x}}{V_{\delta x}} dx\right) \ .$$

4.5.3 No Selection

Recall that in the case of no mutation and no selection, $M_{\delta x} = 0$ and $V_{\delta x} = \frac{x(1-x)}{2N}$. This implies that $G(x) = 1$ and $u(p) = p$. This is the exact result we derived in a different way above, which states that the probability of ultimate fixation of a neutral mutation is given simply by its frequency.

4.5.4 Genic Selection

In the case of genic selection, $h = 0.5$ and the fitnesses of the individual genotypes are $\{1, 1+s, 1+2s\}$. Letting x be the frequency of the selected allele,

$$M_{\delta x} = \frac{x^2(1+2s) + (1+s)x(1-x)}{\bar{w}} - x = \frac{sx(1-x)}{1+2xs} \, .$$

If s is small, $M_{\delta x} \approx sx(1-x)$, $G(x) = \exp(-4Nsx)$, and $u(p \mid s) = \frac{1-\exp(-4Nsp)}{1-\exp(-4Ns)}$. This implies that the probability of fixation of a new mutation is

$$u\left(\frac{1}{2N} \mid s\right) = \frac{1-\mathrm{e}^{-2s}}{1-\mathrm{e}^{-4Ns}} \approx \frac{2s}{1-\mathrm{e}^{-4Ns}}$$

using the fact that $\mathrm{e}^x \approx 1 + x$ if x is small.

Since the mutation process for both selected and neutral mutations is Poisson, their relative substitution rates are given by the ratio of the probabilities of fixation *assuming independence among sites*. Let ω equal the ratio of the probability of fixation of a selected mutation per selected site relative to the probability of fixation of a neutral mutation per neutral site:

$$\omega = \frac{f_s u(p \mid s \neq 0)}{f_0 u(p \mid s = 0)} = \frac{f_s}{f_0} \frac{\frac{2s}{1-\mathrm{e}^{-4Ns}}}{\frac{1}{2N}} = \frac{f_s}{f_0} \frac{2\gamma}{1-\mathrm{e}^{-2\gamma}} \, .$$

As previously mentioned, ω can be interpreted as the expected dn/ds ratio assuming silent mutations are neutral. We will assume $f_0 = f_s$ for the remainder of the chapter (for coding DNA). As we see from Figures 4.1, 4.3, and 4.5, even modest amounts of natural selection can have a profoundly strong effect on rates of substitution. For example, it has been estimated that the historical effective population size of humans is close to $N = 10^5$ (for a review, see [106]). This implies that sites where a mutation would lower the expected number of offspring an individual contributes to the next generation by as little as 0.0025% ($\gamma = -5$) would not evolve at any appreciable rate ($\omega < 0.01$).

In the case of positive genic selection, as s becomes large, the probability of ultimate fixation for a new mutation is well-approximated by $u \approx 2s$ and the expected ratio of substitution rates for selected to neutral mutations by $\omega \approx 2\gamma$. This implies that if mutations at some class of sites increased the expected number of offspring by as little as 0.0025% ($\gamma = 5$), they would evolve at 10 times the rate of neutral mutations.

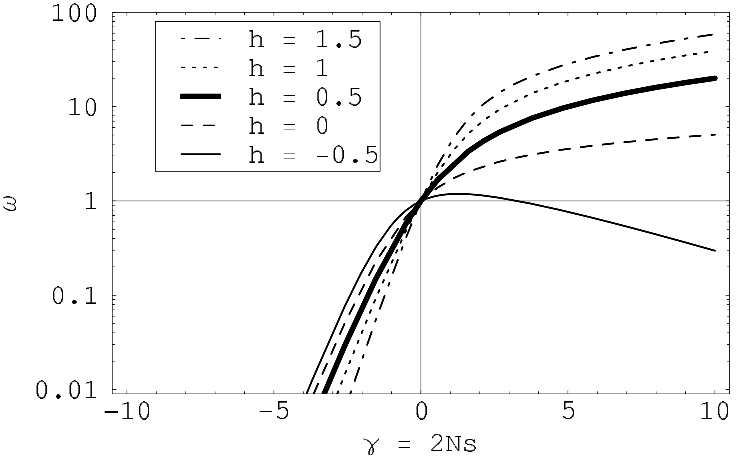


Fig. 4.3. Effect of dominance and selection on rates of molecular evolution.

4.5.5 Dominance

In the case of general selection, it follows directly from the Wright-Fisher model that $M_{\delta x} \approx s(h + (1 - 2h)x)x(1 - x)$ if s is small. This implies that $G(x) = \exp(-4Nshx + 2Ns(1 - 2h)x^2)$ and

$$u(p) = \frac{\int_0^p \mathrm{e}^{-2\gamma shx+\gamma(1-2h)x^2} dx}{\int_0^1 \mathrm{e}^{-2\gamma shx+\gamma(1-2h)x^2} dx} \; .$$

This integral can be evaluated numerically to investigate the effect of heterozygous fitness on rates of molecular evolution. As we see from Figure 4.3, the most profound effects occur when mutations are selectively favored ($\gamma > 0$) and produce heterozygote advantage ($h > 1$). This condition is known as overdominance and such a mutation is said to be subject to balancing selection. In an infinitely large population, overdominance leads to a stable equilibrium in gene frequency such that both alleles are maintained in the population indefinitely. In a finite population, though, higher heterozygote fitness translates into a higher substitution rate relative to neutrality as well as relative to genic selection ($h = 0.5$). The reason for these perplexing results is that having a high heterozygote fitness decreases the probability that a mutation will be lost from the population and thus increases the probability that it will ultimately become fixed in the population.

Another interesting case to consider is that of a mutation whose fitness relative to the wildtype depends on whether it is in heterozygous or homozygous form ($h = -0.50$). If the mutation is deleterious in homozygous form but advantageous in heterozygous form, the mutation will have a slightly higher rate of fixation relative to the case when the heterozygote has intermediate fitness ($h = 0.5$). Alternatively, a beneficial mutation in homozygous form that produces heterozygotes that are less fit than either homozygote will have a lower substitution rate. In interpreting these results, it is important to remember that in estimating ω we are assuming independence among sites. As we will see below, linkage among selected sites can cause interference effects that will counter the single-site dynamics illustrated in Figure 4.3. This is particularly true in the case of strong dominance.

4.6 Variation in Selection Among Sites

Understanding how the distribution of selection coefficients among newly arising mutations affects the rates and patterns of molecular evolution has been a focus of extensive research in theoretical population genetics. In a series of papers, Tomoko Ohta (along with Kimura) [81, 76, 77, 78] first investigated the molecular evolution of "nearly neutral" mutations and found that their behavior was quite different from that of strictly neutral mutations ($\gamma = 0$). In particular, she showed that if there is a high rate of input of slightly deleterious mutations ($-2 < \gamma < 0$) into a population, then this class of mutations can contribute significantly to the overall substitution rate even though these mutations are slightly less fit than the existing wildtype allele. As discussed in Section 4.2, Ohta and Kimura also demonstrated that a nearly neutral model would predict a negative correlation between population size and rate of molecular evolution since natural selection is more efficient in a larger population.

The original work of Ohta and Kimura went on to inspire a plethora of nearly neutral, nonneutral, and fluctuating-environment population genetics theories of molecular evolution. For example, Ohta proposed the exponential-shift model [79], where $-s$ follows an exponential distribution among new mutations (the term shift is used since s is relative to the wildtype allele and the distribution must shift after an allele fixes in the population). Likewise, Kimura [59] suggested a Gamma-shift model that conveniently had sufficient mass near $s = 0$ to account for several neutral and nearly neutral predictions [61, 31]. Ohta and Tachida [82] also proposed a fixed fitness model, where the distribution of s was Gaussian and independent of parental type (a so-called house-of-cards model). These models have been used to argue that if a substantial proportion of slightly deleterious mutations are input into the population, the rate of fixation contributes significantly to the proportion of mutations that fix in the population. It is important to note, though, that the conclusion comes directly from assumptions regarding the functional form of the distribution of selective effects among sites. Since there is no biological reason to favor one distribution over another a priori, in practical applications it is important to be catholic on the matter and consider several potential candidate distributions.

Recently, two methods have come on the market for estimating the distribution of selective effects among new mutations. Nielsen and Yang [73] have developed a likelihood-based method for use with divergence data that considers ten different models (e.g., constant, normal, Gamma, exponential, normal + invariant). (A similar method was suggested by Felsenstein [25] but to our knowledge not fully implemented.) Nielsen and Yang applied their model to a data set of eight mtDNA primate genomes and found that of the models considered, a normal or Gamma-shift model with some sites held invariant was the best fit to data (and significantly better than an exponential distribution [79]). Likewise, Stanley Sawyer and colleagues have developed a method for fitting a normal-shift model to polymorphism and divergence data [90] and applied it to 56 loci with polymorphism from *Drosophila simulans* and divergence data relative to a *D. melanogaster* reference strain. In these models, it is assumed that selection coefficients at a given site are constant in time and do not depend on the nucleotide present. Below we present a brief analysis of the normal-shift model and discuss the findings of Nielsen and Yang [73] and Sawyer et al. [90] in light of the analysis.

4.6.1 Normal Shift

Assume that we starts a Poisson number of Wright-Fisher processes at rate $2N\mu$ per generation and that these processes do not interfere with one another. The number of processes that fix for the selected mutation in some interval of time t will be Poisson with rate

$$\begin{aligned}\mathbb{E}(K \mid \gamma) &= 2N\mu t u(s)\\ &= \mu t \frac{2\gamma}{1-\mathrm{e}^{-2\gamma}}\\ &= \mu t k(\gamma).\end{aligned}$$

Likewise, if mutations have a distribution of selection coefficients such that the probability that a mutation has selection coefficient γ is governed by $f(\gamma)$, then the number of mutations that fix will be Poisson with rate

$$\mathbb{E}(K) = \mu t \int_{-\infty}^{\infty} k(\gamma) f(\gamma) d\gamma \ . \tag{4.18}$$

We can now calculate some statistics of interest. For example, the distribution of selection coefficients among fixed mutations (f is for “fixed”) is

$$p_f(\gamma) = \frac{k(\gamma) f(\gamma) d\gamma}{\int_{-\infty}^{\infty} k(\gamma) f(\gamma) d\gamma} \ . \tag{4.19}$$

This implies that the average selection coefficient of substitutions can be easily computed as

$$\mathbb{E}_f(\gamma) = \int_{-\infty}^{\infty} \gamma p_f(\gamma) d\gamma \ . \tag{4.20}$$

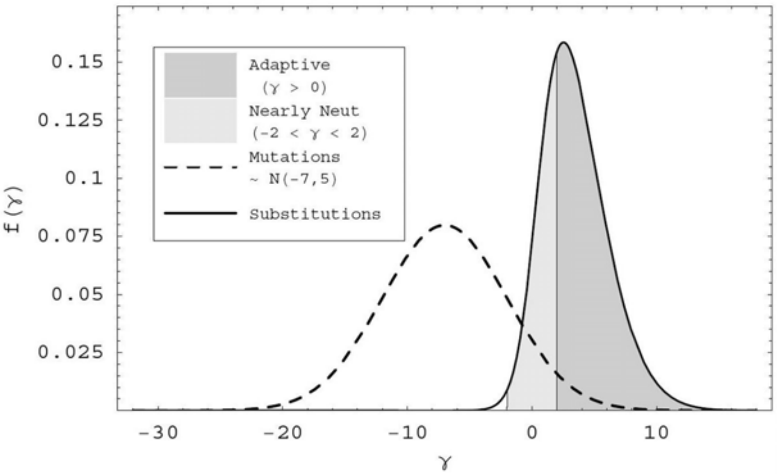


Fig. 4.4. A Gaussian model ("normal shift") for the distribution of selection coefficients among mutations [90]. In this example, the selection coefficient of new mutations is normally distributed with mean $\mu = -7$ and standard deviation $\sigma = 5$. In this example, 68.9% of substitutions are adaptive (dark grey area), 30.7% are nearly neutral, and 0.4% are deleterious.

Likewise, the proportion of fixed differences that are nearly neutral (using the definition of nearly neutral as $-2 \leq \gamma \leq 2$) is

$$p_f(-2 \leq \gamma \leq 2) = \frac{\int_{-2}^{2} k(\gamma) f(\gamma) d\gamma}{\int_{-\infty}^{\infty} k(\gamma) f(\gamma) d\gamma} \tag{4.21}$$

and the proportion of fixed differences that are positively selected (and not nearly neutral) is given by the tail probability

$$p_f(\gamma > 2 \mid \zeta) = \frac{\int_{2}^{\infty} k(\gamma) f(\gamma) d\gamma}{\int_{-\infty}^{\infty} k(\gamma) f(\gamma) d\gamma} \,. \tag{4.22}$$

In Figures 4.4 and 4.5, we explore the effects of a Gaussian model for the distribution of selection coefficients among newly arising mutations. In Figure 4.4, mutations are assumed to follow a normal distribution with mean $\mu = -7$ and standard deviation $\sigma = 5$. Using (4.21) and (4.22), we can estimate the proportion of substitutions that are nearly neutral and adaptive via standard numerical integration (grey areas under the solid curve in Figure 4.4). We note that in this example the vast majority of mutations are deleterious (> 91% are below 0), while most of the substitutions (fixed differences) are positively selected: 92.8% are above $\gamma = 0$, and 68.3% have a selection coefficient above $\gamma = 2$. The average selection coefficient of fixed mutations is a (surprisingly) high $\gamma = 3.49$.

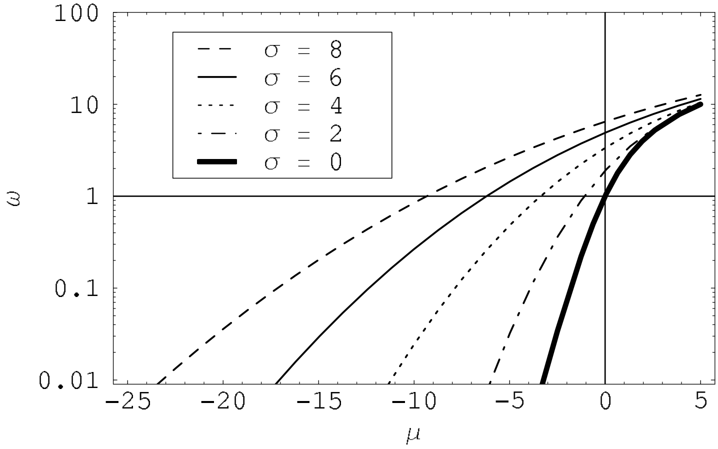


Fig. 4.5. Effect of variance in the distribution of selection coefficients among newly arising mutations on rates of molecular evolution. In this figure, μ is the mean of the distribution of selective effects.

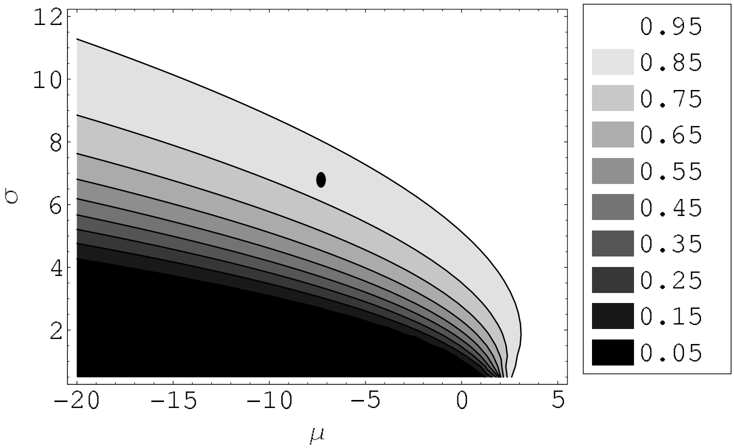


Fig. 4.6. Proportion of adaptive substitutions ($s > \frac{1}{N}$) as a function of the mean of the distribution of selection coefficients for new mutations μ and standard deviation σ. The black point represents the estimated mean and variance for a typical *Drosophila* gene [90].

The fact that mutations differ in their selective effects also has a strong implication for interpreting the ω ratio. In Figure 4.5, we plot the expected ω ratio for varying levels of selection (where the x-axis is the average selected effect of the new mutation) and variability among mutations assuming $f_s = f_0$, where σ corresponds to the standard deviation of selection coefficients among new mutations. In the case of moderate variance $\sigma = 6$, as long as the average selective effect of newly arising mutations is greater than -5, the ω ratio will

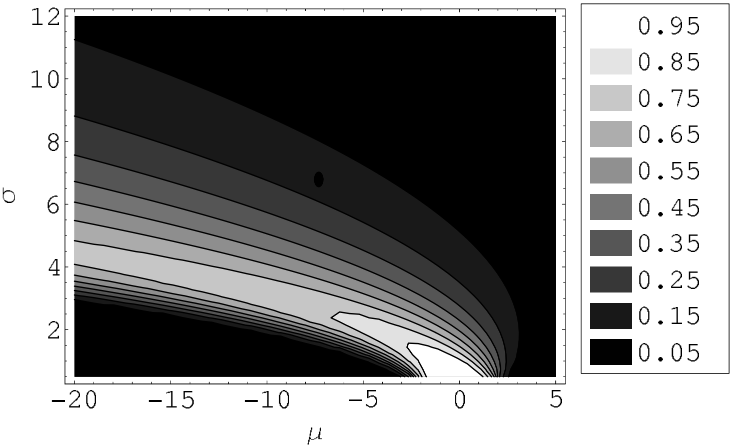


Fig. 4.7. Proportion of nearly neutral substitutions ($|s| \leq \frac{1}{N}$) as a function of the mean of the distribution of selection coefficients for new mutations μ and standard deviation σ. The black point represents the estimated mean and variance for a typical *Drosophila* gene [90].

be greater than 1 (even though most mutations are deleterious). This explains a perplexing phenomenon that is observed in day-to-day analysis of DNA sequence evolution: namely, how it is that one can detect positive selection in the first place if most of the amino acid sites in a protein are rather constrained. The answer is that natural selection is extremely efficient at fixing even slightly favored mutations, so that as long as there is some reasonable fraction of mutations that are adaptive, the average rate of fixation for selected sites (e.g., amino acid sites) may outstrip the neutral rate of evolution. In Figure 4.6, we plot the proportion of fixed differences that are adaptive as a function of both the average selective effect of new mutations (μ) and standard deviation (σ). We note that as long as the standard deviation among newly arising mutations is greater than 6, most of the substitutions will be adaptive even if, on average, mutations are extremely deleterious. The comparable contour plot for nearly neutral mutations is given in Figure 4.7. These simple results bolster the idea that comparing the rate of substitution for different types of sites in protein-coding genes is an effective way of detecting positively selected sites.

The results of Sawyer et al. [90] bear a strong resemblance to the pattern we have just described. They estimated the distribution of selective effects among new mutations in a typical *Drosophila* gene to have mean $\mu = -7.31$ and $\sigma = 6.79$. This implies that close to 97.1% of amino substitutions in a typical *Drosophila* nuclear gene are of positively selected mutations ($\gamma > 0$), with 84.7% being clearly adaptive, $\gamma \geq 2$; see (4.22). Furthermore, close to 15.2% of substitutions are of "nearly neutral" mutations ($-2 \leq \gamma \leq 2$), with only 2.7% being "slightly deleterious" ($-2 \leq \gamma \leq 0$) mutations while 12.4% are "slightly advantageous" ($0 \leq \gamma \leq 2$). Lastly, the average selection coefficient of substituting mutations is 5.67; see (4.20). The black disks in Figures 4.6 and 4.7 correspond to the Sawyer et al. estimate of μ and σ for *Drosophila.* These results are consistent with previous findings of adaptive protein evolution in *Drosophila* (e.g., [95, 24, 8, 84]).

4.6.2 Linkage

One interpretation of the normal-shift model is that of "Darwin's wedge" at a molecular level [90]. As Darwin wrote in *The Origin of Species* [20, cp. 3]

> In looking at Nature, it is most necessary to keep the foregoing considerations always in mind never to forget that every single organic being around us may be said to be striving to the utmost to increase in numbers... . The face of Nature may be compared to a yielding surface, with ten thousand sharp wedges packed close together and driven inwards by incessant blows, sometimes one wedge being struck, and then another with greater force.

In this passage, Darwin views natural selection as competition for fixed resources leading to rapid turnover of species. That is, one wedge forces another out in order to fix its claim to a space in a cramped environment. At a molecular level, the metaphor works well: a slightly favored mutation sweeping through the population acts as a wedge to displace the existing alleles at a given locus. The efficacy of such a wedging scheme, of course, is predicated upon the frequency of favored wedges. If there are too many favored mutations competing for fixation at a given locus, they will knock each other out of competition and the efficacy of selection can be greatly reduced. In many ways, the fact that one can detect positive selection in the face of interference among selected sites is in fact *stronger* evidence for a selective model of molecular evolution. That is to say, if one estimates that the average selection coefficient of fixed mutations is $\gamma = 5.67$ in the presence of interference, the true selection coefficient on the mutation must be higher. There is relatively strong support for the view that linkage can affect rates and patterns of substitution for selected mutations [7, 42, 15, 36, 37].

For example, Birky and Walsh [7] have shown analytically and via simulation that linked selected mutation negatively interferes so as to increase the rate of substitution of deleterious mutations and to decrease the rate of substitution of advantageous mutations. They attribute this phenomenon to a reduction in the effective population size through an increase in the variance of offspring among individuals. As we saw in Section 4.5, if the effective population size of a species is reduced, genetic drift begins to play a more prominent role in determining the evolutionary fate of mutations.

The predictions of interference selection hypotheses have gained strong support in recent years. For example, Comeron and Kreitman used analytical, simulation, and genomic analyses to demonstrate that interference selection can explain patterns of codon usage and intron size in *Drosophila* [15]. Likewise, a prediction of the interference hypothesis is that rates of adaptive evolution should be reduced in regions of low recombination since the tighter the linkage among favored mutations, the stronger the interference effects. There is experimental evidence that regions of low recombination in *Drosophila* do, in fact, show a reduction in the rate of adaptive evolution [93, 5], as do non-recombining mitochondria [111, 84]. Likewise, if we consider the analysis of Nielsen and Yang [73], they estimate a distribution of selective effects among mutations in primate mtDNA that has mean $\mu = -1.72$ and $\sigma = 0.72$. For such a model, the proportion of substitutions that have a selection coefficient greater than $\gamma = 0$ is a quite small 6%, consistent with the view that linkage limits the rate of adaptive evolution.

There is also important literature on the impact of linkage on rates of evolution in nearly neutral and fluctuating selection models [32, 33, 34, 35, 17, 18]. Much of it has focused on analytical and simulation work for describing which population genetics models lead to an overdispersed molecular clock. To summarize all of this work, Gillespie and Cutler have shown that the overdispersed molecular clock cannot readily be explained by overdominance, underdominance, a rapidly fluctuating environment, or the nearly neutral models presented above (although certain narrow parameter ranges can lead to an over-dispersed clock, the models do not, in general, lead to an overdispersed clock). Gillespie [32] has found that a slowly fluctuating environment can lead to an over-dispersed clock if the oscillations are on the same order as the mutation rate. Likewise, Cutler [17] has argued that a simple deleterious model that shifts between a favored and a deleterious allele is sufficient to explain the overdispersed clock.

Gillespie has also investigated the effects of linkage and selection on the relationship between population size and the rate of molecular evolution using extensive simulations. He has identified three domains, which he terms the Darwin domain ($k_s \propto N$), the Kimura domain ($k_s \approx \mu f_0$), and the Ohta domain ($k_s \propto \frac{1}{N}$). Not surprisingly, he finds that the nearly neutral models (exponential shift [79], Gamma-shift model [59], and house of cards [82]) all fall within the Ohta domain where the rate of evolution is inversely proportional to population size. He also finds that the normal-shift model with mean $\mu = 0$ (Darwin's wedge) appropriately falls in the Darwin domain, where the rate of substitution is proportional to the population size. He also notes that the rate of substitution for the normal-shift model is substantially reduced relative to the expectation under the independence-among-sites model (4.18) (as one might predict from [7]). Lastly, he finds, surprisingly, that the fluctuating selection, neutral, and overdominance models all lead to the Kimura domain, where the rate of molecular evolution is independent of the population size. A mechanism that Gillespie has proposed to explain this last observation is the theory of genetic drift, whereby positive selection on one locus leads to the reduction of effective population size at linked neutral loci even in an infinitely large population [36, 37].

Lastly, Brian Charlesworth and colleagues have also shown that linkage of neutral mutations to deleterious mutations ("background" selection) [9, 10, 11] leads to a chronic and pronounced reduction in the local effective population size of a chromosomal region. Recent experimental work on patterns of variation within the nonrecombining neo-sex chromosomes of *Drosophila miranda* [2, 3] has confirmed some theoretical predictions of the background selection model. Likewise, Cutler [18] has argued that the background selection hypothesis is consistent with the observed overdispersed molecular clock.

Acknowledgments

This work was supported by grants NSF-0319553 and USDA-HATCH(NYC-151411). I am indebted to Rasmus Nielsen and Ryan Hernandez for comments and suggestions on earlier drafts. Ryan also helped produce Figure 4.1. Likewise, Scott Williamson provided assistance in producing Figure 4.3. I would also like to thank the Institute for Pure and Applied Mathematics (IPAM) for administrative support and office space where some of this work was completed. Beatrix Johnson is also acknowledged for help in library research.

References

[1] J. M. Akey, G. Zhang, K. Zhang, L. Jin, and M. D. Shriver. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res*, 12(12):1805–1814, Dec 2002.

[2] D. Bachtrog. Adaptation shapes patterns of genome evolution on sexual and asexual chromosomes in *Drosophila. Nat Genet*, 34(2):215–219, Jun 2003.

[3] D. Bachtrog. Protein evolution and codon usage bias on the neo-sex chromosomes of *Drosophila miranda. Genetics*, 165(3):1221–1232, Nov 2003.

[4] M. Barrier, C. D. Bustamante, J. Yu, and M. D. Purugganan. Selection on rapidly evolving proteins in the *Arabidopsis* genome. *Genetics*, 163(2):723–733, Feb 2003.

[5] A. J. Betancourt and D. C. Presgraves. Linkage limits the power of natural selection in *Drosophila. Proc Natl Acad Sci USA*, 99(21):13616–13620, Oct 2002.

[6] N. Bierne and A. Eyre-Walker. The genomic rate of adaptive amino-acid substitution in *Drosophila. Mol Biol Evol*, Mar 2004.

[7] C. W. Birky and J. B. Walsh. Effects of linkage on rates of molecular evolution. *Proc Natl Acad Sci USA*, 85(17):6414–6418, Sep 1988.

[8] C. D. Bustamante, R. Nielsen, S. A. Sawyer, K. M. Olsen, M. D. Purugganan, and D. L. Hartl. The cost of inbreeding in *Arabidopsis. Nature*, 416(6880):531–534, Apr 2002.

[9] B. Charlesworth. The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genet Res*, 63(3):213–227, Jun 1994.

[10] B. Charlesworth, M. T. Morgan, and D. Charlesworth. The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134(4):1289–1303, Aug 1993.

[11] D. Charlesworth, B. Charlesworth, and M. T. Morgan. The pattern of neutral molecular variation under the background selection model. *Genetics*, 141(4):1619–1632, Dec 1995.

[12] M. Choisy, C. H. Woelk, J. F. Guegan, and D. L. Robertson. Comparative study of adaptive molecular evolution in different human immunodeficiency virus groups and subtypes. *J Virol*, 78(4):1962–1970, Feb 2004.

[13] A. G. Clark, S. Glanowski, R. Nielsen, P. D. Thomas, A. Kejariwal, M. A. Todd, D. M. Tanenbaum, D. Civello, F. Lu, B. Murphy, S. Ferriera, G. Wang, X. Zheng, T. J. White, J. J. Sninsky, M. D. Adams, and M. Cargill. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science*, 302(5652):1960–1963, Dec 2003.

[14] R. M. Clark, E. Linton, J. Messing, and J. F. Doebley. Pattern of diversity in the genomic region near the maize domestication gene tb1. *Proc Natl Acad Sci USA*, 101(3):700–707, Jan 2004.

[15] J. M. Comeron and M. Kreitman. Population, evolutionary and genomic consequences of interference selection. *Genetics*, 161(1):389–410, May 2002.

[16] D. J. Cutler. Clustered mutations have no effect on the overdispersed molecular clock: A response to Huai and Woodruff. *Genetics*, 149(1):463–464, May 1998.

[17] D. J. Cutler. The index of dispersion of molecular evolution: Slow fluctuations. *Theor Popul Biol*, 57(2):177–186, Mar 2000.

[18] D. J. Cutler. Understanding the overdispersed molecular clock. *Genetics*, 154(3):1403–1417, Mar 2000.

[19] J. da Silva. The evolutionary adaptation of HIV-1 to specific immunity. *Curr HIV Res*, 1(3):363–371, Jul 2003.

[20] C. Darwin. *The Origin of Species.* Oxford University Press, Oxford, reissue edition, 1859.

[21] W. J. Ewens. A note on the sampling theory for infinite alleles and infinite sites models. *Theor Popul Biol*, 6(2):143–148, Oct 1974.

[22] W. J. Ewens. A note on the variance of the number of loci having a given gene frequency. *Genetics*, 80(1):221–222, May 1975.

[23] W. J. Ewens. *Mathematical Population Genetics.* Springer, New York, 2004.

[24] J. C. Fay, G. J. Wyckoff, and C. I. Wu. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. *Nature*, 415(6875):1024–1026, Feb 2002.

[25] J. Felsenstein. Taking variation of evolutionary rates between sites into account in inferring phylogenies. *J Mol Evol*, 53(4–5):447–455, Oct 2001.

[26] J. Felsenstein and G. A. Churchill. A hidden Markov model approach to variation among sites in rate of evolution. *Mol Biol Evol*, 13(1):93–104, Jan 1996.

[27] R. A. Fisher. On the dominance ratio. *Proc Roy Soc Edinburgh*, 42:321–341, 1922.

[28] R. A. Fisher. *The Genetical Theory of Natural Selection.* Clarendon Press, Oxford, 1st edition, 1930.

[29] Y. Gilad, C. D. Bustamante, D. Lancet, and S. Paabo. Natural selection on the olfactory receptor gene family in humans and chimpanzees. *Am J Hum Genet*, 73(3):489–501, Sep 2003.

[30] J. H. Gillespie. A general model to account for enzyme variation in natural populations. v. the sas-cff model. *Theor Popul Biol*, 14:1–45, 1978.

[31] J. H. Gillespie. *The Causes of Molecular Evolution.* Oxford University Press, Oxford, 1991.

[32] J. H. Gillespie. Substitution processes in molecular evolution. I. Uniform and clustered substitutions in a haploid model. *Genetics*, 134:971–981, 1993.

[33] J. H. Gillespie. Substitution processes in molecular evolution. II. Exchangeable models from population genetics. *Evolution*, 48:1101–1113, 1994.

[34] J. H. Gillespie. Substitution processes in molecular evolution. III. Deleterious alleles. *Genetics*, 138:943–952, 1994.

[35] J. H. Gillespie. The role of population size in molecular evolution. *Theor Popul Biol*, 55:145–156, 1999.

[36] J. H. Gillespie. Genetic drift in an infinite population. The pseudohitchhiking model. *Genetics*, 155(2):909–919, Jun 2000.

[37] J. H. Gillespie. The neutral theory in an infinite population. *Gene*, 261(1):11–18, Dec 2000.

[38] S. Glinka, L. Ometto, S. Mousset, W. Stephan, and D. De Lorenzo. Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: A multi-locus approach. *Genetics*, 165(3):1269–1278, Nov 2003.

[39] N. Goldman and Z. Yang. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol*, 11(5):725–736, Sep 1994.

[40] M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*, 22(2):160–174, 1985.

[41] I. Hellmann, S. Zollner, W. Enard, I. Ebersberger, B. Nickel, and S. Paabo. Selection on human genes as revealed by comparisons to chimpanzee cDNA. *Genome Res*, 13(5):831–837, May 2003.

[42] W. G. Hill and A. Robertson. The effect of linkage on limits to artificial selection. *Genet Res*, 8(3):269–294, Dec 1966.

[43] J. L. Hubby and R. C. Lewontin. A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in *Drosophila pseudoobscura*. *Genetics*, 54(2):577–594, Aug 1966.

[44] R. R. Hudson. Gene genealogies and the coalescent process. In *Oxford Surveys in Evolutionary Biology 7*, pages 1–44. Oxford University Press, Oxford, 1990.

[45] J. P. Huelsenbeck and R. Nielsen. Effect of nonindependent substitution on phylogenetic accuracy. *Syst Biol*, 48(2):317–328, Jun 1999.

[46] G. A. Huttley, M. W. Smith, M. Carrington, and S. J. O'Brien. A scan for linkage disequilibrium across the human genome. *Genetics*, 152(4):1711–1722, Aug 1999.

[47] V. Jaenicke-Despres, E. S. Buckler, B. D. Smith, M. T. Gilbert, A. Cooper, J. Doebley, and S. Paabo. Early allelic selection in maize as revealed by ancient DNA. *Science*, 302(5648):1206–1208, Nov 2003.

[48] T. H. Jukes and C. R. Cantor. Evolution of protein molecules. In *Mammalian Protein Metabolism*, pages 21–132. Academic Press, New York, 1969.

[49] S. Karlin and H. M. Taylor. *A Second Course in Stochastic Processes.* Academic Press, New York, 1981.

[50] M. Kayser, S. Brauer, and M. Stoneking. A genome scan to detect candidate regions influenced by local natural selection in human populations. *Mol Biol Evol*, 20(6):893–900, Jun 2003.

[51] F. P. Kelly. *Reversibility and Stochastic Networks.* John Wiley & Sons, Chichester, 1979.

[52] M. Kimura. Solution of a process of random genetic drift with a continuous model. *Proc Natl Acad Sci USA*, 41:114–150, 1955.

[53] M. Kimura. On the probability of fixation of mutant genes in a population. *Genetics*, 47:713–719, 1962.

[54] M. Kimura. Diffusion models in population genetics. *J Appl Probab*, 1:177–232, 1964.

[55] M. Kimura. Evolutionary rate at the molecular level. *Nature*, 217:624–626, 1968.

[56] M. Kimura. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61(4):893–903, Apr 1969.

[57] M. Kimura. The rate of molecular evolution considered from the standpoint of poulation genetics. *Proc Natl Acad Sci USA*, 63:1181–1188, 1969.

[58] M. Kimura. Theoretical foundation of population genetics at the molecular level. *Theor Popul Biol*, 2(2):174–208, Jun 1971.

[59] M. Kimura. Models of effectively neutral mutations in which selective constraint is incorporated. *Proc Nat Acad Sci USA*, 76:3440–3444, 1979.

[60] M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*, 16(2):111–120, Dec 1980.

[61] M. Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Camebridge, 1983.

[62] M. Kimura and T. Ohta. Protein polymorphism as a phase of molecular evolution. *Nature*, 229(5285):467–469, Feb 1971.

[63] J. L. King and T. H. Jukes. Non-Darwinian evolution. *Science*, 164:788–798, 1969.

[64] R. C. Lewontin. *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York, 1974.

[65] R. C. Lewontin and J. L. Hubby. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics*, 54(2):595–609, Aug 1966.

[66] W. H. Li. Distribution of nucleotide differences between two randomly chosen cistrons in a finite population. *Genetics*, 85(2):331–337, Feb 1977.

[67] T. Mitchell-Olds and M. J. Clauss. Plant evolutionary genomics. *Curr Opin Plant Biol*, 5(1):74–79, Feb 2002.

[68] C. B. Moore, M. John, I. R. James, F. T. Christiansen, C. S. Witt, and S. A. Mallal. Evidence of HIV-1 adaptation to HLA-restricted immune responses at a population level. *Science*, 296(5572):1439–1443, May 2002.

[69] S. V. Muse. Evolutionary analyses of DNA sequences subject to constraints of secondary structure. *Genetics*, 139(3):1429–1439, Mar 1995.

[70] S. V. Muse and B. S. Gaut. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol*, 11(5):715–724, Sep 1994.

[71] R. Nielsen and J. Wakeley. Distinguishing migration from isolation: A Markov chain Monte Carlo approach. *Genetics*, 158(2):885–896, Jun 2001.

[72] R. Nielsen and Z. Yang. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, 148(3):929–936, Mar 1998.

[73] R. Nielsen and Z. Yang. Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. *Mol Biol Evol*, 20(8):1231–1239, Aug 2003.

[74] J. R. Norris. *Markov Chains*. Cambridge University Press, Cambridge, 1997.

[75] D. Nurminsky, D. D. Aguiar, C. D. Bustamante, and D. L. Hartl. Chromosomal effects of rapid gene evolution in *Drosophila melanogaster*. *Science*, 291(5501):128–130, Jan 2001.

[76] T. Ohta. Evolutionary rate of cistrons and DNA divergence. *J Mol Evol*, 1:150–157, 1972.

[77] T. Ohta. Slightly deleterious mutant substitutions in evolution. *Nature*, 246:96–98, 1973.

[78] T. Ohta. Mutational pressure as the main cause of molecular evolution and polymorphism. *Nature*, 252:351–354, 1974.

[79] T. Ohta. Extension of the neutral mutation drift hypothesis. In M. Kimura, editor, *Molecular Evolution and Polymorphism*, pages 148–167. National Institute of Genetics, Mishima, Japan, 1977.

[80] T. Ohta. Near-neutrality in evolution of genes and gene regulation. *Proc Natl Acad Sci USA*, 99(25):16134–16137, Dec 2002.

[81] T. Ohta and M. Kimura. On the constancy of the evolutionary rate of cistrons. *J Mol Evol*, 1:18–25, 1971.

[82] T. Ohta and H. Tachida. Theoretical study of nearly neutrality. I. Heterozygosity and rate of mutant substitution. *Genetics*, 126:219–229, 1990.

[83] B. A. Payseur, A. D. Cutter, and M. W. Nachman. Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. *Mol Biol Evol*, 19(7):1143–1153, Jul 2002.

[84] D. M. Rand, D. M. Weinreich, and B. O. Cezairliyan. Neutrality tests of conservative-radical amino acid changes in nuclear- and mitochondrially-encoded proteins. *Gene*, 261(1):115–125, Dec 2000.

[85] D. M. Robinson, D. T. Jones, H. Kishino, N. Goldman, and J. L. Thorne. Protein evolution with dependence among codons due to tertiary structure. *Mol Biol Evol*, 20(10):1692–1704, Oct 2003.

[86] N. A. Rosenberg. The probability of topological concordance of gene trees and species trees. *Theor Popul Biol*, 61(2):225–247, Mar 2002.

[87] P. C. Sabeti, D. E. Reich, J. M. Higgins, H. Z. Levine, D. J. Richter, S. F. Schaffner, S. B. Gabriel, J. V. Platko, N. J. Patterson, G. J. McDonald, H. C. Ackerman, S. J. Campbell, D. Altshuler, R. Cooper, D. Kwiatkowski, R. Ward, and E. S. Lander. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909):832–837, Oct 2002.

[88] S. A. Sawyer. On the past history of an allele now known to have frequency p. *J Appl Probab*, 14:439–450, 1977.

[89] S. A. Sawyer and D. L. Hartl. Population genetics of polymorphism and divergence. *Genetics*, 132(4):1161–1176, Dec 1992.

[90] S. A. Sawyer, R. J. Kulathinal, C. D. Bustamante, and D. L. Hartl. Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection. *J Mol Evol*, 57 (Suppl 1):S154–S164, 2003.

[91] K. J. Schmid, L. Nigro, C. F. Aquadro, and D. Tautz. Large number of replacement polymorphisms in rapidly evolving genes of *Drosophila*: Implications for genome-wide surveys of DNA polymorphism. *Genetics*, 153(4):1717–1729, Dec 1999.

[92] M. Schoniger and A. von Haeseler. A stochastic model for the evolution of autocorrelated DNA sequences. *Mol Phylogenet Evol*, 3(3):240–247, Sep 1994.

[93] L. A. Sheldahl, D. M. Weinreich, and D. M. Rand. Recombination, dominance and selection on amino acid polymorphism in the *Drosophila* genome: Contrasting patterns on the x and fourth chromosomes. *Genetics*, 165(3):1195–1208, Nov 2003.

[94] A. Siepel and D. Haussler. Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol Biol Evol*, 21(3):468–488, Mar 2004.

[95] N. G. Smith and A. Eyre-Walker. Adaptive protein evolution in *Drosophila*. *Nature*, 415(6875):1022–1024, Feb 2002.

[96] J. F. Storz and M. W. Nachman. Natural selection on protein polymorphism in the rodent genus *Peromyscus*: Evidence from interlocus contrasts. *Evol Int J Org Evol*, 57(11):2628–2635, Nov 2003.

[97] S. Sunyaev, F. A. Kondrashov, P. Bork, and V. Ramensky. Impact of selection, mutation rate and genetic drift on human genetic variation. *Hum Mol Genet*, 12(24):3325–3330, Dec 2003.

[98] W. J. Swanson, A. G. Clark, H. M. Waldrip-Dail, M. F. Wolfner, and C. F. Aquadro. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. *Proc Natl Acad Sci USA*, 98(13):7375–7379, Jun 2001.

[99] N. Takahata. *Population Genetics, Molecular Evolution, and the Neutral Theory*. University of Chicago Press, Chicago, 1994.

[100] N. Takahata, K. Ishii, and H. Matsuda. Effects of temporal fluctuation of selection coefficient on gene frequency in a population. *Proc Natl Acad Sci USA*, 72:4541–4545, 1975.

[101] N. Takahata and M. Nei. Gene genealogy and variance of interpopulational nucleotide differences. *Genetics*, 110(2):325–344, Jun 1985.

[102] N. Takahata, Y. Satta, and J. Klein. Divergence time and population size in the lineage leading to modern humans. *Theor Popul Biol*, 48(2):198–221, Oct 1995.

[103] M. I. Tenaillon, M. C. Sawkins, L. K. Anderson, S. M. Stack, J. Doebley, and B. S. Gaut. Patterns of diversity and recombination along chromosome 1 of maize (*Zea mays* L. ssp.). *Genetics*, 162(3):1401–1413, Nov 2002.

[104] J. L. Thorne. Models of protein sequence evolution and their applications. *Curr Opin Genet Dev*, 10(6):602–605, Dec 2000.

[105] J. Wakeley and J. Hey. Estimating ancestral population parameters. *Genetics*, 145(3):847–855, Mar 1997.

[106] J. D. Wall. Estimating ancestral population sizes and divergence times. *Genetics*, 163(1):395–404, Jan 2003.

[107] G. A. Watterson. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol*, 7(2):256–276, Apr 1975.

[108] G. A. Watterson. Mutant substitutions at linked nucleotide sites. *Adv Appl Probab*, 14:206–224, 1982.

[109] G. A. Watterson. Substitution times for a mutant nucleotide. *J Appl Probab*, 19A:59–70, 1984.

[110] C. Weinig, L. A. Dorn, N. C. Kane, Z. M. German, S. S. Halldorsdottir, M. C. Ungerer, Y. Toyonaga, T. F. Mackay, M. D. Purugganan, and J. Schmitt. Heterogeneous selection at specific loci in natural environments in *Arabidopsis thaliana*. *Genetics*, 165(1):321–329, Sep 2003.

[111] D. M. Weinreich. The rates of molecular evolution in rodent and primate mitochondrial DNA. *J Mol Evol*, 52(1):40–50, Jan 2001.

[112] D. M. Weinreich and D. M. Rand. Contrasting patterns of nonneutral evolution in proteins encoded in nuclear and mitochondrial genomes. *Genetics*, 156(1):385–399, Sep 2000.

[113] S. Whelan, P. Lio, and N. Goldman. Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends Genet*, 17(5):262–272, May 2001.

[114] D. E. Wildman, M. Uddin, G. Liu, L. I. Grossman, and M. Goodman. Implications of natural selection in shaping 99.4% nonsynonymous DNA identity between humans and chimpanzees: Enlarging genus *Homo*. *Proc Natl Acad Sci USA*, 100(12):7181–7188, Jun 2003.

[115] S. Williamson. Adaptation in the env gene of HIV-1 and evolutionary theories of disease progression. *Mol Biol Evol*, 20(8):1318–1325, Aug 2003.

[116] S. Wright. Evolution in mendelian populations. *Genetics*, 160:97–159, 1931.

[117] S. Wright. The distribution of gene frequencies under irreversible mutation. *Proc Natl Acad Sci USA*, 24:253–259, 1938.

[118] W. Yang, J. P. Bielawski, and Z. Yang. Widespread adaptive evolution in the human immunodeficiency virus type 1 genome. *J Mol Evol*, 57(2):212–221, Aug 2003.

[119] Z. Yang. Estimating the pattern of nucleotide substitution. *J Mol Evol*, 39(1):105–111, Jul 1994.

[120] Z. Yang. A space-time process model for the evolution of DNA sequences. *Genetics*, 139(2):993–1005, Feb 1995.

[121] Z. Yang. Maximum likelihood analysis of adaptive evolution in HIV-1 gp120 env gene. In *Pacific Symposium on Biocomputing*, pages 226–237. World Scientific, Singapore, 2001.

[122] Z. Yang. Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics*, 162(4):1811–1823, Dec 2002.

[123] Z. Yang and R. Nielsen. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol*, 19(6):908–917, Jun 2002.

[124] Z. Yang, R. Nielsen, N. Goldman, and A. M. Pedersen. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, 155(1):431–449, May 2000.

[125] E. Yuste, A. Moya, and C. Lopez-Galindez. Frequency-dependent selection in human immunodeficiency virus type 1. *J Gen Virol*, 83(Pt 1):103–106, Jan 2002.

[126] E. Zuckerkandl and L. Pauling. Evolutionary divergence and convergence in proteins. In V. Bryson and H. Voge, editors, *Evolving Genes and Proteins*, pages 97–166. Academic Press, New York, 1965.

Part II

Practical Approaches for Data Analysis