Chapter 5

Statistical Methods for Detecting the Presence of Natural Selection in Bacterial Populations

Yun-Xin Fu and Xiaoming Liu

5.1 INTRODUCTION

Natural selection is one of the most powerful mechanisms determining the fate of a population and thus, elucidating the impact of natural selection has always been an important aspect of population study. Natural selection in the evolution of a population often leaves traces at the molecular level. Therefore, samples from a population or from multiple related populations or species can be used to reveal the signature of natural selection. Many times, various sampling strategies are required to detect the presence of natural selection at different evolutionary time scales. Molecular sequences sampled from sufficiently divergent populations or species are needed for dissecting natural selection that persists for long periods of time. In such situations, many nucleotide sites have accumulated multiple mutations so that the relative tendency of nucleotide changes can be evaluated, which provide the basis for judging if natural selection is an important evolutionary force and if so, the type of natural selection. A large body of literature exists in this area (e.g., Nielsen and Yang, 1998; Suzuki and Gojobori, 1999; Yang and Nielsen, 2002).

When one is interested in the evolutionary forces that govern the recent significant events of a population, including natural selection that is operating on the extant population, samples from individuals within the sample populations or within closely related populations are necessary. One main characteristic of such samples is that most of the observed polymorphic sites have experienced few mutations or just one mutation. When studying a pathogen population, it is often necessary to take multiple samples from the evolving population over a period of time so that molecular changes can be tracked. To reveal the presence of natural selection from a sample within a population and from longitudinal samples, a different statistical approach than those used when studying long-term evolution is needed.

This chapter focuses on statistical methods for detecting the presence of recent natural selection that can be revealed from within population samples. The review is not meant to

Bacterial Population Genetics in Infectious Disease, Edited by D. Ashley Robinson, Daniel Falush, and Edward J. Feil

be comprehensive as we are more interested in statistical tests that are applicable to bacterial population studies. We will start with the general predictions of the outcome of natural selection, then describe a few widely used statistical methods and end with the discussion of some statistical approaches that are specific to the study of bacterial populations.

5.2 NATURAL SELECTION

Due to the haploidy of a bacterial genome, the type of natural selection in a bacterial population is limited and the outcome is relatively easy to predict: The better allele will ultimately win. A straightforward classical demonstration of the prediction is as follows.

Consider two alleles, A and a, at a locus of a bacterial population, with fitness W_A and W_a, respectively. Suppose the frequency of the two alleles are p_{t-1} and q_{t-1} at generation $t-1$, then in generation t at reproduction time,

$$p_t = \frac{p_{t-1}W_A}{p_{t-1}W_A + q_{t-1}W_a}. \tag{5.1}$$

Therefore, the ratio of p_t and q_t is

$$\frac{p_{t-1}W_A}{q_{t-1}W_a} = \ldots = \left(\frac{p_0}{q_0}\right)\left(\frac{W_A}{W_a}\right)^t. \tag{5.2}$$

Suppose A is the fitter allele ($W_A > W_a$), then the ratio will approach infinity as t goes to infinity, which indicates that A will be fixed in the population eventually.

The above demonstration assumes that population size is sufficiently large so that the formula for p_t is accurate. Since typically bacterial population size is large (at least census size), it is traditionally thought that random drift is not a significant factor. This may be appropriate when dealing with a situation in which the new allele is sufficiently advantageous over the existing one; however, in reality, it is often difficult to identify an advantageous allele from a snapshot of the population in the form of a sample of DNA sequences, in which many polymorphic sites are present. Although collectively a bacterial species is usually large indeed, its population is often geographically structured, and selection may proceed differently in different local populations in which random genetic drift can become a significant factor. The observation of many mutations of various frequencies in a sample of DNA sequences of reasonable length from a bacterial population is a strong indication that random genetic drift is important in dealing with the recent evolution of bacteria.

When the locus under study is subject to purifying selection, that is, some mutations are deleterious, the prediction by Equation 5.2 is that their frequencies should decrease each generation when the population size is infinitely large. However, random genetic drift has played a significant role in keeping some deleterious mutations in the population longer, and the signature of such mutations can be found by the pattern of polymorphism in a sample.

5.3 STATISTICAL METHODS FOR DETECTING THE PRESENCE OF NATURAL SELECTION

5.3.1 Summary Statistics of Polymorphism

There has been a long history in both biological sciences as well as in the statistical field to gain insight into a scientific query through the use of quantities that summarize impor-

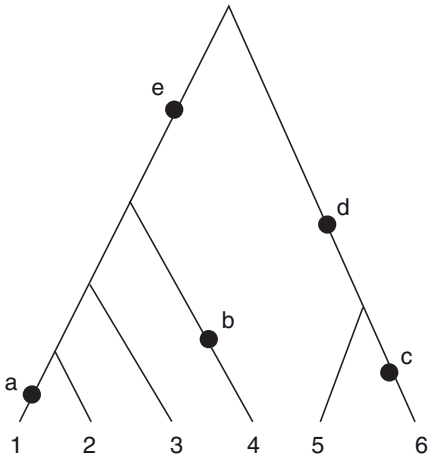

Figure 5.1 A genealogy of six sequences with five mutations (a–e) since the most recent common ancestor, three of which (a, b, and c) are of size 1, one of which (d) is of size 2, and one of which (e) is of size 4.

tant features of the data. These quantities are commonly known as summary statistics. The polymorphism in a sample of DNA sequences from a population can be summarized by a number of summary statistics from which a number of statistical tests have been developed. Three most widely known summary statistics are the number of distinct alleles (k), the number of segregating sites (K), and the mean number of nucleotide differences between two sequences in a sample (Π). A mutation observed in a sample must have occurred in the genealogy of the sample and can be further classified into size classes, which are the number of sequences that carry the mutant nucleotide. For a sample of n sequences, a mutation is thus of size from 1 to $n - 1$. Most summary statistics can be expressed as linear functions of the number of mutations of various classes (ξ_i, $i = 1, \ldots, n - 1$). To illustrate the definitions of various summary statistics, consider the following hypothetical sample of six sequences of 15 bps:

1: GAGGCTCTGATCCCA
2: AAGGCTCTGATCCCA
3: AAGGCTCTGATCCCA
4: AAGGTTCTGATCCCA
5: AAGGCTCTGATCTCG
6: AAGGCTCAGATCTCG

which resulted from the genealogy in Fig. 5.1. By direct counting, it is found that $k = 5$, and from the genealogy, it follows that $\xi_1 = 3$, $\xi_2 = 1$, and $\xi_4 = 1$, while $\xi_2 = \xi_2 = 0$. The number of segregating sites K is equal to five, which is also the number of mutations in the genealogy. Let d_{ij} represent the number of nucleotide differences between sequences i and j, then it is easy to see, for example, $d_{12} = d_{13} = 1$, $d_{14} = 2$, and the average of all the d_{ij} leads to $\Pi = 2.07$.

Under the infinite allele model, that is, every mutation in the population creates a new allele, the number of distinct alleles, k, in a sample of n sequences has the following distribution (Ewens, 1972; Karlin and McGregor, 1972):

$$Pr(k|\theta) = \frac{|S_k|\theta^k}{S_n(\theta)}, \tag{5.3}$$

where $S_n(\theta) = \theta(\theta - 1). \ldots (\theta - n + 1)$ and S_k is the coefficient of θ^k when $S_n(\theta)$ is expanded to the polynomial of θ, also known as the Stirling number of the first kind (Abramowitz

and Stegun, 1965). Furthermore, if n_i is the occurrence of allele type $i(i = 1, \ldots, k)$, we have

$$Pr(n_1, n_2, \ldots, n_k | k) = \frac{n!}{|S_n^k| k! n_1 n_2 \ldots n_k}. \tag{5.4}$$

Under the infinite site model (i.e., every mutation occurs in a new site),

$$K = \xi_1 + \ldots + \xi_{n-1}, \tag{5.5}$$

which in this example leads to $K = 3 + 1 + 0 + 1 + 0 = 5$. By definition, $\Pi = \frac{2}{n(n-1)} \sum_{i<j} d_{ij}$. When there is no recombination, Π can be expressed as

$$\Pi = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} i(n-i)\xi_i. \tag{5.6}$$

In this example, we can compute Π from the above formula, which results in Π equal to 2.07. Many potentially useful linear functions of $\xi_i(i = 1, \ldots, n-1)$ can be defined, for example,

$$m = \frac{1}{n-1}(\xi_1 + 2\xi_2 + \cdots + (n-1)\xi_{n-1}), \tag{5.7}$$

which is the mean number (or more appropriately corrected) of mutations in a sequence since the most recent common ancestor (MRCA). Another interesting quantity is

$$h = \frac{2}{n(n-1)}\left(\xi_1 + 2^2\xi_2 + \ldots + (n-1)^2\xi_{n-1}\right), \tag{5.8}$$

which places a heavier weight on mutations of large sizes. For the example, $m = 1.8$ and $h = 1.53$.

When the sample is taken from a population evolving according to the Wright–Fisher model with constant population size, and all mutations are selectively neutral (the so-called neutral Wright–Fisher model), it follows (Fu, 1995) that $E(\xi_i) = \frac{\theta}{i}$, where $E(\xi_i)$ represents the mathematical expectation of ξ_i, from which it is easy to show that

$$E(K) = a_n\theta, \tag{5.9}$$

$$E(\Pi) = \theta, \tag{5.10}$$

$$E(m) = \theta, \text{ and} \tag{5.11}$$

$$E(h) = \theta, \tag{5.12}$$

where $a_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n-1}$. Historically, $E(K)$ and $E(\Pi)$ (as well as their variances) were first derived without referring to ξ_i by Watterson (1975) and Tajima (1983), respectively. The variances of these summary statistics are as follows:

$$\mathrm{Var}(K) = a_n\theta + b_n\theta^2 \tag{5.13}$$

$$\mathrm{Var}(\Pi) = \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9(n(n-1))}\theta^2, \tag{5.14}$$

where $b_n = 1 + \frac{1}{2^2} + \ldots + \frac{1}{(n-1)^2}$ and Var stands for variance. The variances of all the summary statistics described above as well as their covariance can be derived from those of $\xi_i(i = 1, \ldots, n-1)$. The case of ξ_1 is of special interest since ξ_1 is a widely used quantity. Fu and Li (1993) shows that

$$\mathrm{Var}(\xi_1) = \theta + 2\left[na_n - \frac{2(n-1)}{(n-1)(n-2)}\right]\theta^2, \tag{5.15}$$

while from Fu (2009) (also see Zeng et al., 2006), we have

$$Var(m) = \frac{n}{2(n-1)}\theta + \left(2\left(\frac{n}{n-1}\right)^2(b_{n+1}-1)-1\right)\theta^2. \tag{5.16}$$

Although it is more powerful to use summary statistics that are linear functions of $\xi_i (i = 1, \ldots, n-1)$, sometimes the values of ξ_i may be difficult to identify. In such cases, it is preferable to use $\eta_i = \xi_i + \xi_{n-1}$ $(i = 1, \ldots, n/2)$ as the building blocks for summary statistics. Two such summary statistics are K and Π. The variance and covariance between each pair of η_i are given by Fu (1995). For bacterial and viral population studies, often longitudinal samples are taken. In addition to the summary statistic described above, there are some new informative quantities. One such quantity is the number of private (unique) mutations to a sample taken at a specific time.

5.3.2 Statistical Test of Neutrality Based on Summary Statistics

In the presence of natural selection, for example, some of the mutations in the sample are deleterious, or some are advantageous, or the sequences are from a locus that is linked to another one, which is the target of natural selection (genetic hitchhiking). The expectation is that almost all the summary statistics described in the previous section will be changed, but the extent of change varies from situation to situation, and more importantly from statistic to statistic.

When there are some deleterious mutations in the locus being sequenced, it is expected that most of these deleterious mutations are either quickly removed from the population or are kept in low frequencies. Therefore, summary statistics that are influenced strongly by low-frequency mutations are expected to be inflated in the presence of deleterious mutations. To be more specific, in the presence of many deleterious mutations, the number of singleton mutations (ξ_1) will be high. Therefore, summary statistics that weigh heavily on ξ_1, such as ξ_1, will be inflated severely. The number of segregating sites (K) is also expected to be inflated, even more pronounced when compared with Π, which gives much less weight on mutations of low frequencies.

A similar effect to most summary statistics will be observed if the sample is taken from a locus that has experienced (or tightly linked to one) a recent fixation of an advantageous allele. This is because when an advantageous allele reaches fixation from an initial low frequency, it mimics a population whose size is expanding relatively fast; since ξ_1 represents mutations that are on average young in age, a large population size corresponds to a large value for ξ_1. Similarly, K is expected to be inflated more severely than the value of Π. However, if the sample is taken at a time before fixation is completed, the sequences in the sample may fall into two types: one carries the advantageous alleles (or linked to), and another does not carry the advantageous allele (or linked to). If the advantageous allele is nearly fixed, the class that carries the advantageous allele will be in high frequency, and the mutations that separate the two classes of sequences will be in relatively high frequency as well. Summary statistics that weigh heavily on high-frequency mutant classes will be inflated; among the summary statistics described above, h, as defined by Equation 5.8, is one such statistic.

It should be pointed out that altered expectations of summary statistics are not always due to the presence of natural selection. As we have mentioned above, a rapidly growing population can lead to inflated numbers of low-frequency mutants. A structured population, on the other hand, will lead to the presence of an excess of mutations of intermediate frequencies. Therefore, summary statistics such as Π are likely inflated in the presence of population structure.

The responses of summary statistics to departure from neutrality (i.e., evolving according to the Wright–Fisher model with constant population size and all mutations are selectively neutral) lead to a class of statistical tests that is of the form

$$\frac{L_1 - L_2}{\sqrt{Var(L_1 - L_2)}} \tag{5.17}$$

where $E(L_1) = E(L_2) = \theta$ under neutrality, but are likely to be different in the presence of natural selection. Therefore, a significant departure from zero is taken as evidence against neutrality. The reason for the denominator is to standardize the test statistic so that it is not affected by or at least not sensitive to unknown values of θ. Note that

$$\mathrm{Var}(L_1 - L_2) = \mathrm{Var}(L_1) + \mathrm{Var}(L_2) - 2\mathrm{Cov}(L_1, L_2), \tag{5.18}$$

where $\mathrm{Cov}(L_1, L_2)$ stands for the covariance between L_1 and L_2.

Therefore, as long as the variance of two summary statistics and their covariance are known, the variance of their difference can be computed. Even with the standardization, such a statistic does not usually follow some standard distribution. Therefore, its critical values are normally determined from simulated samples, which can be performed easily using efficient coalescent algorithms.

The first such statistical test was proposed by Tajima (1989) and was known as Tajima's D test, which has $L_1 = \Pi$ and $L_2 = \frac{K}{a_n}$. The covariance between L_1 and L_2 is found by Tajima (1989) as

$$\mathrm{Cov}(L_1, L_2) = \frac{\theta}{a_n} + \frac{n+2}{2na_n}\theta^2. \tag{5.19}$$

To compute the value of the variance, an estimate of θ is required. In the case of Tajima's test, θ is estimated by $\frac{K}{a_n}$, which is known as Watterson's estimator (Watterson, 1975), and θ^2 is estimated by $\frac{K(K-1)}{a_n^2 + b_n}$.

Fu and Li (1993) proposed several tests utilizing rare mutants. Their D test correspondsto $L_1 = \frac{K}{a_n}$ and $L_2 = \xi_1$. For this test, the covariance between L_1 and L_2 is

$$\mathrm{Cov}\left(\frac{K}{a_n}, \xi_1\right) = \theta + \frac{a_n}{n-1}\theta^2. \tag{5.20}$$

Again, to evaluate the variance, θ is estimated by Watterson's estimator. Fay and Wu (2000) proposed a test using $L_1 = \Pi$ and $L_2 = h$, but it turns out that this is equivalent to a test with $L_1 = \Pi$ and $L_2 = m$. The covariance between Π and m is given by Fu (2009) (also, see Zeng et al., 2006) as

$$\mathrm{Cov}(\Pi, m) = \frac{n+1}{3(n-1)}\theta + \frac{7n^2 + 3n - 2 - 4n(n+1)b_{n+1}}{2(n-1)^2}\theta^2. \tag{5.21}$$

Another line of statistical tests is to utilize Ewens' sampling formula (Ewens, 1972; Karlin and McGregor, 1972). The first well-known test of this type is Watterson's (1978)

homozygosity test, which was motivated by that conditional on the number of alleles in a sample, the frequencies of each allele are independent of population parameter θ. The test statistic is as follows:

$$H = \sum_i \left(\frac{n_i}{n} \right)^2. \tag{5.22}$$

Although Watterson's test is appropriate when detailed DNA sequence variation is not available, it is in general less powerful when compared with statistical tests that utilize such detailed patterns of DNA variation. One way to utilize Ewens' sampling formulas is to compare the number of distinct alleles in a sample with a predicted number under certain assumptions. Given the value of θ, too many k and too few k can be taken as evidence against neutrality. Fu (1997) proposed to substitute θ by Π, and this led to the following test:

$$F_S = \log\left(\frac{s}{1-s} \right), \tag{5.23}$$

where $s = \Sigma_{i=1}^{k} Pr(k|\theta = \Pi)$ and $Pr(k|\theta = \Pi)$ is computed by Equation 5.3. This test is found to be particularly powerful for detecting the access of rare alleles; it is also sensitive to the presence of recombination.

Most of the statistical tests described above as well as the test to be described in the next section can be found in a number of popular softwares used for analyzing population data. These include DnaSP (Librado and Rozas, 2009), Arlequin (Excoffier et al., 2005), and NeutralityTest (Li and Fu, 2009).

5.3.3 Statistical Test Utilizing Both within and between Population Variations

The statistical tests described in the previous section utilize only within-sample polymorphism; often, samples from multiple closely related species are available, and it is desirable to utilize interspecific variation as well. There are two well-known tests of the kind known as the MK test (McDonald and Kreitman, 1991) and the HKA test (Hudson et al., 1987). We shall describe the MK test in this section.

Consider two closely related populations from each of which a sample of DNA sequences of a protein-coding region is taken. In the total sample, a polymorphic site may be such that all the sequences in one sample possess one particular nucleotide, while all the sequences in the other sample possess another different nucleotide. This type of polymorphism is called between-sample variation; otherwise, it is called within-sample variation. Since the sequences are from a protein-coding region, each polymorphism will be either a synonymous change or a nonsynonymous change. The pattern of polymorphism in the total sample can thus be summarized in a 2 × 2 table:

	Within sample	Between sample
Synonymous	a	b
Nonsynonymous	c	d

where a, for example, is the number of polymorphic sites that are both within-sample variation and synonymous change. When mutations are selectively neutral, it is expected that the ratio of nonsynonymous and synonymous changes (dN/dS) remains constant over

time. That is, under neutrality, $\frac{a}{c} = \frac{b}{d}$. This equality can be tested statistically by a chi-square test, which results in the following test statistic:

$$X^2 = \frac{n(ad - bc)^2}{[(a+b)(a+c)(b+d)(c+d)]} \tag{5.24}$$

where $n = a + b + c + d$ is the total number of polymorphic sites. When n is sufficiently large, X^2 can be approximated by a X^2 variable with one degree of freedom (df). Significantly large values of X^2 are taken as evidence against neutrality. Alternatively, the G test or Fisher's exact test can be used when n is small. Note that such a test can easily be extended to more than two species.

Typically, a significant departure in the MK test is caused by an excess of nonsynonymous between-sample variation, which is taken as evidence of positive selection in favor of some amino acid changes. Since MK is a widely applicable and powerful test, its validity has been a subject of debate. Early debate partially stemmed from confusion of terminology. A discussion can be found in Fu (2000). Eyre-Walker (2002) found that existence of some deleterious mutations and increasing population sizes can lead to a significant MK test. Rocha et al. (2006) suggested that for closely related populations, dN/dS depends on their separation time and that a lag in the removal of slightly deleterious mutations may explain the change of dN/dS over time. Therefore, caution is also needed for inferences of selection based on the MK test.

5.4 STATISTICAL METHODS FOR BACTERIAL POPULATIONS

5.4.1 Longitudinal Samples

Longitudinal samples are samples taken at different time points from the same population (Fig. 5.2). For genetic studies of most organisms with relatively low mutation rates, longitudinal samples can be pooled together as a single sample taken at the same time, which simplifies the analysis. The justification of such convention is that the sampling interval is so small that the possible mutations accumulated on the sequences studied within the sampling intervals are negligible. However, for fast-evolving organisms, including some bacteria, some sampling intervals (in years) may be sufficiently long to allow for the observation of significant genetic change within samples. Although new statistical methods have been developed for analyzing longitudinal DNA samples (see review by Drummond et al., 2003), few methods are available for detecting the presence of natural selection. On the other hand, if longitudinal samples can be safely pooled as a single sample, more methods for selection detection can be applied (see Sections 5.3.2 and 5.3.3). For longitudinal samples taken from fast-evolving bacterial populations, it is suggested to test whether there are significant genetic changes between longitudinal samples as the first step. If not, the samples can be pooled as a single sample and methods for selection detection for single samples can be used. Otherwise, the methods designed for longitudinal samples should be applied.

Liu and Fu (2007) proposed several methods for testing genetical isochronism or for detecting significant genetical heterochronism in longitudinal samples. Here we introduce a test based on the number of private mutations within samples. Suppose there are two samples taken from an evolving haploid population at time t_0 and $t_0 + t$, respectively, where t is the sampling interval in generations. Let n_1 and n_2 be the sizes of samples taken at t_0

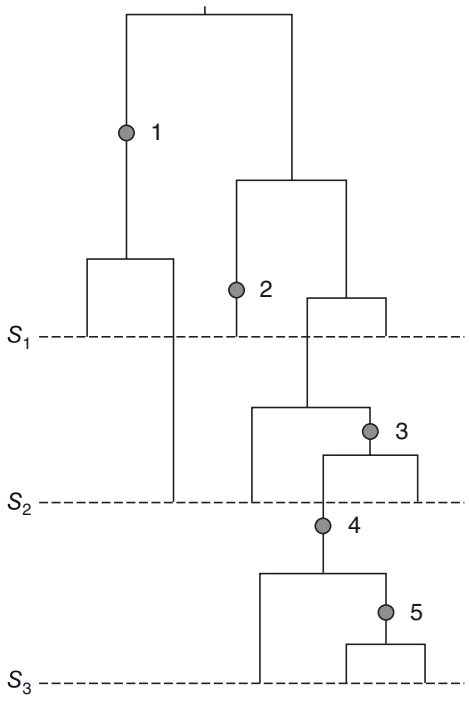

Figure 5.2 A genealogy of three longitudinal samples, s_1, s_2, and s_3, sampled at three different time points. Each sample has three sequences.

and $t_0 + t$, respectively. The number of private mutations within a sample is the number of sites that are not only polymorphic in that sample but are monomorphic in the other samples. Let $K_p(i)(i = 1, 2)$ be the number or private mutations of sample i. Then the test statistic is

$$T_c = \frac{c\left(K_p(1) - E\left(K_p(1)\right)\right) + (1-c)\left(K_p(2) - E\left(K_p(2)\right)\right)}{\sqrt{\mathrm{Var}\left(cK_p(1) + (1-c)K_p(2)\right)}}. \tag{5.25}$$

The detailed computation of mean and variance can be found in Liu and Fu (2007), in which several values of c were compared using simulation. It was found that the test statistic with $c_2 = \frac{n_2}{n_1 + n_2}$ or $c_3 = \frac{n_2^2}{n_1^2 + n_2^2}$ has the highest power. The significance level of the test can be determined by either permutation or coalescent simulation.

If significant genetical heterochronism between longitudinal samples is suggested by the test, then the samples should not be pooled and analyzed as a single sample. Unfortunately, there are only a very few studies on detecting potential selection with longitudinal samples.

Goode et al. (2008) extended Nielsen and Yang's (1998) codon model for protein-coding sequences to apply to longitudinal samples. Using their model, nucleic sites can be assigned to different selection categories (negatively selected, positively selected, and neutral). However, their method is based on an inferred phylogenetic tree and does not take into account the uncertainty of phylogeny reconstruction.

Edwards et al. (2006) and later Drummond and Suchard (2008) tried to overcome this shortcoming and to take into account the uncertainty of the gene genealogy and the parameters of the mutation model and the demographic model at the same time. To do so, their methods sample gene genealogies of the sequences, along with model parameters using a Markov chain Monte Carlo (MCMC) framework (Drummond et al., 2002). More specifically, the genealogy and parameters are sampled according to the posterior probability

distribution $Pr(G, \Omega, \theta|Y)$, where G is the gene genealogy, θ is the mutation parameters, and Ω is the parameter for an exponential growth model. Given each sampled genealogy, six different statistics are calculated. They include one classic neutrality test statistic (Fu and Li's [1993] D-statistic), two measures of branch length distribution (age of the MRCA and total tree length), and three measures of tree imbalance (Kirkpatrick and Slatkin's [1993] B_1, McKenzie and Steel's [2000] C_n, and Colless's [1982] I_c). With a large number of genealogies sampled, empirical null distributions of these statistics can be obtained. If the same statistics observed in the original sample are significantly unlikely to be observed from the null distributions, the null hypothesis of neutrality is rejected. This method is theoretically promising, although several caveats need to be considered. First, it assumes no recombination in the sequences and so far, it is unknown how robust the method is when recombination cannot be ignored. Second, a large parameter space needs to be explored, so the power and reliability of the conclusion may be a concern if the sample size is not large. Third, there are always some technical issues for the application of MCMC, such as prior choices and convergence detection.

5.4.2 Selection Based on DNA Fingerprints

In the area of studying bacterial populations, DNA fingerprints are often used to determine the polymorphic level or diversity of the population. Typically, bacteria (or clones, or strains) were sampled from a population. Then, a specified locus was amplified using polymerase chain reaction (PCR) for each bacterium. Finally, certain DNA fingerprinting methods (such as PCR-SSCP and PCR-DGGE) were used to identify the alleles of the locus (Nocker et al., 2007). If each bacterium can be independently genotyped using DNA fingerprinting, the frequencies of different alleles $n_i (i = 1, \ldots, k)$ can be directly counted. Then, Watterson's homozygosity test (Equation 5.22) can be applied.

Sometimes, the smallest sampling unit may still consist of multiple bacteria. For example, a host was infected by multiple strains of pathogens, and a sample taken from that host may contain more than one strain. Targeting this problem, Rannala et al. (2000) used a Poisson distribution to model the number of strains in each sample and a multinomial distribution to model the number of allele copies carried by these strains. Based on this model, they proposed a maximum likelihood estimator of the allele frequency given the observed presence/absence frequency of each allele. Anderson and Scheet (2001) derived another estimator of allele frequency from the same model, and their estimator is supposed to be less biased when compared to Rannala et al.'s (2000) original estimator. After the allele frequency is estimated, a similar test based on Ewens' sampling formulas can be conducted. However, it is not clear to what extent uncertainty in the allelic frequencies will affect the neutrality test.

5.4.3 Selection Based on the Presence/Absence of Certain Genomic Islands (GIs)

One important mechanism of bacterial genome evolution is horizontal gene transfer. Many of the accessory genes transferred by this mechanism form a distinct DNA segment called GI (Juhas et al., 2009). Studies have shown that GIs may be associated with many important adaptive functions, such as pathogenicity, symbiosis, sucrose, aromatic compound metabolism, mercury resistance, and siderophore synthesis (Juhas et al., 2009). However, since GIs typically carry various novel genes (no detectable homologues in other species)

(Hsiao et al., 2005), a statistical test of association between a GI and a phenotype is needed to detect adaptive GIs. Instead of testing the correlation between the concentration of the presence/absence of one binary character with the presence/absence of another binary character, as Maddison (1990)'s concentrated changes test, here we propose a test of nonrandom copresence/coabsence of two binary characters (one GI, one phenotype, or two GIs) on branches of a given phylogeny. It is possible to further extend the method by taking gene genealogy uncertainty into account, as Edwards et al. (2006) and Drummond and Suchard (2008) did. We assume that in addition to the presence/absence data of GIs of interest, each strain in the sample is also assayed by other markers, such as multilocus sequence typing (MLST).

First, a phylogenetic tree is reconstructed using MLST. Then, given the presence/absence states of the GIs on the external nodes of the tree and the phylogeny, the presence/absence states of the internal nodes of the tree (ancestors of the sample) can be inferred using available phylogeny reconstruction programs, such as PAUP* (Swofford, 2003; http://paup.csit.fsu.edu/), PHYLIP (Felsenstein, 1989; http://evolution.genetics.washington.edu/phylip/), or PAML (Yang, 2007; http://abacus.gene.ucl.ac.uk/software/paml.html). Similarly, the presence/absence of certain adaptive phenotypes of the internal nodes can also be inferred. After the states of the GI or phenotype on internal nodes are inferred, the number of state change (i.e., presence to absence or absence to presence) events on the branches of the phylogeny can be counted.

Then some statistical measures of the correlation between the (inferred) phenotype and the (inferred) GI states are calculated, such as Gini impurity (Breiman et al., 1984). For example, if we use 0 and 1 to represent the absence or the presence of a particular trait (GI or phenotype), then n_{00}, n_{01}, n_{10}, and n_{11} are the counts of nodes that have both traits absent, the first absent and the second present, the first present and the second absent, and both present, respectively. Let $n_0 = n_{00} + n_{01}$, $n_1 = n_{10} + n_{11}$, and $n = n_0 + n_1$. Then, the Gini impurity is calculated as

$$G = \left[1 - \left(\frac{n_{01}+n_{11}}{n}\right)^2 - \left(\frac{n_{00}+n_{10}}{n}\right)^2\right] - \frac{n_1}{n}\left[1 - \left(\frac{n_{11}}{n_1}\right)^2 - \left(\frac{n_{10}}{n_1}\right)^2\right] - \frac{n_0}{n}\left[1 - \left(\frac{n_{01}}{n_0}\right)^2 - \left(\frac{n_{00}}{n_{10}}\right)^2\right]. \tag{5.26}$$

A larger Gini impurity measure means a better correlation. The Gini impurity measure can be calculated for internal nodes only, external nodes only, and all nodes combined on the phylogenetic tree.

To test the significance of correlation between a GI and a phenotype, a Monte Carlo simulation can be used by superimposing the state change events onto the phylogeny while fixing their number according to their inferred counts using original data (see details below). For each replication, Gini impurity measures are compared to those calculated with original data, which are designated as G_0. After a large number of replications, the percent of the replications with a larger Gini impurity measure than G_0 was counted. This is the empirical p value for the significance of correlation between the GI and the phenotype, with the null hypothesis that the presence/absence events of each trait independently occur on the phylogeny.

To reasonably simulate the horizontal transfer of the GI, the superimposing process used in the simulation needs to be carefully designed. The process begins with all nodes having the same states (0/1) as the root. If the root state is "1," an absence (deletion) event is randomly superimposed onto a branch with a probability that equals to its branch length

divided by the total length of the branches with state "1." If the root state is "0," then a presence (insertion) event is superimposed onto the phylogeny. After an event is superimposed onto a branch, all the descendant nodes of the branch change their states accordingly. After an event is superimposed, the probability of whether an insertion or deletion is the next event to superimpose is determined by the relative ratio of the total length of the remaining branches with state "0" and state "1." There are two restrictions of the above process. One is that the number of insertions/deletions to be superimposed needs to be fixed to the number of events as inferred using the original data. If an internal node has undetermined states (due to their equality according to the criteria used in the phylogenetic algorithm, such as maximum parsimony) within each replication, its state is randomly assigned while fixing the total number of event changes. The other restriction is that to superimpose an event, there must be some eligible branches to be superimposed. If we encounter a situation, such as an insertion event that needs to be superimposed because the deletion quotas have already dried up, but there are no eligible remaining branches with state "0," then we have to stop the process and restart from the beginning. So this superimposing process is a trial and error simulation. Another limitation of the process is that it does not allow two events to be superimposed onto the same branch. Further developments addressing these problems are needed.

5.5 AN EXAMPLE

A recent example of applying various statistical tests described in this chapter is given by Zhao and Qin (2007). The data set was originally from a study of the phycoerythrin (*ppe*) gene in two ecotypes of *Prochlorococcus*, which are specifically adapted to high light (HL) or low light (LL) conditions (Steglich et al., 2003). In its original analysis, the authors only did phylogenetic analysis and found the monophyletic origin of the HL and LL sequences. They concluded *ppe* is suitable as a sensitive molecular marker to study *Prochlorococcus* populations. Zhao and Qin (2007) reanalyzed the data and applied multiple methods for selection detection.

Zhao and Qin (2007) applied different intraspecific neutrality tests, including Tajima's (1989) D-statistic and Fu and Li's (1993) D^*- and F^*-statistics on the HL- and LL-*ppeB* locus. They found significant negative values for the tests on HL-*ppeB* ($D = -1.9542$, $p < 0.01$, $D^* = -4.2708$, $p < 0.01$, $F^* = -4.1726$, $p < 0.01$), which suggests an excess of rare variants probably due to directional selection or population bottleneck. As to the LL sequences, Tajima's D showed a marginally significant positive value ($D = 3.4205$, $p < 0.05$), suggesting an excess of intermediate variants possibly due to balancing selection or population subdivision. However, Fu and Li's D^*- and F^*-statistics did not show significant departure from the expectation under neutrality, although their values are also positive. Considering the possibility of mutation rate heterogeneity along sites, they also applied Misawa and Tajima's D^+ test (Misawa and Tajima, 1997) on the same data, which is a modified version of Tajima's D under the finite site model. D^+ also showed significant negative values in the HL-*ppeB* sequences but no significant departure from neutrality in the LL-*ppeB* sequences. A likelihood ratio test for neutrality based on phylogeny (Yang and Nielsen, 2002) was then conducted. The result confirmed the hypothesis of positive selection on the HL-*ppeB* sequences. Besides intraspecific tests, Zhao and Qin (2007) also conducted interspecific neutrality tests, including the MK test (McDonald and Kreitman, 1991) and the likelihood ratio test mentioned above. The MK test showed an excess of nonsynonymous fixed substitutions in the *ppeB* and *ppeA* loci (Fisher's exact test,

$p < 0.001$), which suggests a positive selection on those loci since the divergence of *Prochlorococcus* and *Synechococcus*. This hypothesis was confirmed using the likelihood ratio test. By inferring the selection pressures acting on the *ppeB* loci along with the functional structural information, the authors conclude that HL- and LL-*ppeB* should be under different selective pressures, and positive selection may drive HL-*ppeB* to obtain a new function.

5.6 DISCUSSION AND PERSPECTIVE

The theory and statistical methods for detecting the presence of natural selection using samples from within a population or within closely related populations are reviewed in this chapter, and several new statistical approaches specifically designed for bacterial populations, such as for fast-evolving bacterial pathogens and for GIs, are also presented. While many statistical approaches developed earlier can be applied to bacterial populations, there is also the need for methods that are more specific to microorganisms, including bacterial populations. Longitudinal samples from pathogen populations present some challenges that are not found in the traditional one-sample analysis. New summary statistics as well as new statistical methods for detecting natural selection for longitudinal samples likely will be developed in the future. We note that even for a single sample analysis, there is still considerable room for developing new and useful summary statistics, some perhaps in the form described in Fu (2009).

Bacterial genomes often evolve by acquiring novel and foreign genomic elements or sometimes by eliminating some existing segments; such a mechanism many times creates a pattern of presence/absence of a certain element (GI). How to evaluate the importance of the presence or absence event is not trivial. Although we have presented a method to do so, further analysis of this method as well as developing more powerful methods is desirable.

As far as a statistical approach is concerned, most of the methods described are based on comparisons between two summary statistics. One useful extension is to consider such multiple tests simultaneously (e.g., Innan, 2006; Zeng et al., 2006, 2007). Also note that all such tests of natural selection are based on the comparison of data to the prediction of the null model, which usually assumes a variation of neutrality. Such an approach in many ways is desirable since a neutral model is well accepted as the starting point of the data analysis and can be clearly defined. Because of the nature of such analysis, a significant departure from the null model should be interpreted by noting that natural selection is only, albeit important, one of the possible causes. Other causes include population structure, population growth or shrinkage, and even sampling bias. Biased sampling may be even more pronounced in studying bacterial pathogens since samples are often based on opportunity rather than on design. An alternative statistical approach, such as the maximum likelihood approach or even the Bayesian test (e.g., Drummond and Suchard, 2008), may be desirable when the situation warrants. This is typically true when a particular alternative model of evolution can be identified and justified. Statistical tests that take the specific alternative model into consideration may be more powerful, but one must also be cautious in the interpretation because a number of different evolutionary models may all fit the data adequately.

As far as studying infectious diseases is concerned, it is in an exciting stage since new sequencing technologies (such as pyrosequencing) are capable of generating large amounts of data. However, analyzing such data also presents a considerable challenge

(e.g., Eriksson et al., 2008; Rodrigo et al., 2008), which is not unique to the study of bacterial populations. This is due to the large intrinsic error as well as other uncertainties in the sequencing. Developing statistical tests based on such data will be desirable as part of the effort to meet the challenge.

REFERENCES

Abramowitz, M. and Stegun, I. (1965) *Handbook of Mathematical Functions*. New York, Dover Publications, Inc.

Anderson, E. C. and Scheet, P. A. (2001) Improving the estimation of bacterial allele frequencies. *Genetics* **158(3),** 1383–1386.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984) *Classification and Regression Trees*. Kluwer Academic Publishers, Dordrecht.

Colless, D. H. (1982) Review of "Phylogenetics: The theory and practice of phylogenetic systematics." *Systematic Zoology* **31(1),** 100–104.

Drummond, A. J., Nicholls, G. K., Rodrigo, A. G., and Solomon, W. (2002) Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics* **161(3),** 1307–1320.

Drummond, A. J., Pybus, O. G., Rambaut, A. et al. (2003) Measurably evolving populations. *Trends in Ecology and Evolution* **18,** 481–488.

Drummond, A. and Suchard, M. A. (2008) Fully Bayesian tests of neutrality using genealogical summary statistics. *BMC Genetics* **9(1),** 68.

Edwards, C. T. T., Holmes, E. C., Pybus, O. G. et al. (2006) Evolution of the human immunodeficiency virus envelope gene is dominated by purifying selection. *Genetics* **174(3),** 1441–1453.

Eriksson, N., Pachter, L., Mitsuya, Y. et al. (2008) Viral population estimation using pyrosequencing. *PLoS Computational Biology* **4(5),** e1000074.

Ewens, W. J. (1972) The sampling theory of selectively neutral alleles. *Theoretical Population Biology* **3(1),** 87–112.

Excoffier, L., Laval, G., and Schneider S. (2005) Arlequin (version 3.0): An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online* **1,** 47–50.

Eyre-Walker, A. (2002) Changing effective population size and the McDonald-Kreitman test. *Genetics* **162(4),** 2017–2024.

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

Felsenstein, J. (1989) PHYLIP—Phylogeny inference package (version 3.2). *Cladistics* **5,** 164–166.

Fu, Y. X. (1995) Statistical properties of segregating sites. *Theoretical Population Biology* **48,** 172–197.

Fu, Y. X. (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147**, 915–925.

Fu, Y. X. (2000) Neutrality and selection in molecular evolution: Statistical tests. In *Encyclopedia of Life Sciences* http://www.els.net/ (accessed January 6, 2010)

Fu, Y. X. (2009) Variances and covariances of linear summary statistics of segregating sites (manuscript in preparation).

Fu, Y. X. and Li, W. H. (1993) Statistical tests of neutrality of mutations. *Genetics* **133(3),** 693–709.

Goode, M., Guindon, S., and Rodrigo, A. (2008) Modelling the evolution of protein coding sequences sampled from measurably evolving populations. *Genome Informatics* **21,** 150–164.

Hsiao, W. W. L., Ung, K., Aeschliman, D. et al. (2005) Evidence of a large novel gene pool associated with prokaryotic genomic islands. *PLoS Genetics* **1(5),** e62.

Hudson, R. R., Kreitman, M., and Aguade, M. (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics* **116,** 153–159.

Innan, H. (2006) Modified Hudson-Kreitman-Aguade test and two-dimensional evaluation of neutrality tests. *Genetics* **173,** 1725–1733.

Juhas, M., van der Meer, J. R., Gaillard, M. et al. (2009) Genomic islands: Tools of bacterial horizontal gene transfer and evolution. *FEMS Microbiology Reviews* **33(2),** 376–393.

Karlin, S. and McGregor, J. L. (1972) Addendum to a paper of W. Ewens. *Theoretical Population Biology* **5,** 95–105.

Kirkpatrick, M. and Slatkin, M. (1993) Searching for evolutionary patterns in the shape of a phylogenetic tree. *Evolution* **47(4),** 1171–1181.

Li, H. and Fu, Y. X. (2009) NeutralityTest: A Novel Software for Testing Neutrality. http://xfiles.uth.tmc.edu/xythoswfs/webview/_xy-1789858_1 (manuscript in preparation).

Librado, P. and Rozas, J. (2009) DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, 2009 Apr 3. [Epub ahead of print] doi:10.1093/bioinformatics/btp187

Liu, X. and Fu, Y. X. (2007) Test of genetical isochronism for longitudinal samples of DNA sequences. *Genetics* **176,** 327–342.

Maddison, W. P. (1990) A method for testing the correlated evolution of two binary characters: Are gains or losses concentrated on certain branches of a phylogenetic tree? *Evolution* **44(3),** 539–557.

McDonald, J. H. and Kreitman, M. (1991) Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* **351,** 652–654.

McKenzie, A. and Steel, M. (2000) Distributions of cherries for two models of trees. *Mathematical Biosciences* **164(1),** 81–92.

Misawa, K. and Tajima, F. (1997) Estimation of the amount of DNA polymorphism when the neutral mutation rate varies among sites. *Genetics* **147,** 1959–1964.

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

Nocker, A., Burr, M., and Camper, A. K. (2007) Genotypic microbial community profiling: A critical technical review. *Microbial Ecology* **54,** 276–289.

Rannala, B., Qiu, W. G., and Dykhuizen, D. E. (2000) Methods for estimating gene frequencies and detecting selection in bacterial populations. *Genetics* **155(2),** 499–508.

Rocha, E. P. C., Maynard Smith, J., Hurst, L. D. et al. (2006) Comparisons of dN/dS are time dependent for closely related bacterial genomes. *Journal of Theoretical Biology* **239,** 226–235.

Rodrigo, A., Bertels, F., Heled, J. et al. (2008) The perils of plenty: What are we going to do with all these genes? *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* **363**(1512**),** 3893–3902.

Steglich, C., Post, A. F., and Hess, W. R. (2003) Analysis of natural populations of *Prochlorococcus* spp. in the northern Red Sea using phycoerythrin gene sequences. *Environmental Microbiology* **5,** 681–690.

Suzuki, Y. and Gojobori T. (1999) A method for detecting positive selection at single amino acid sites. *Molecular Biology and Evolution* **16,** 1315–1328.

Swofford, D. L. (2003) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods) Version 4. http://paup.csit.fsu.edu/ (accessed January 5, 2010).

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

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

Watterson, G. A. (1975) On the number of segregating sites. *Theoretical Population Biology* **7,** 256–276.

Watterson, G. A. (1978) The homozygosity test of neutrality. *Genetics* **88(2),** 405–417.

Yang, Z. (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24(8),** 1586–1591

Yang, Z. and Nielsen, R. (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution* **19,** 908–917.

Zeng, K., Fu, Y. X., Shi, S., and Wu, C-I. (2006) Statistical tests for detecting positive selection by utilizing high frequency variants. *Genetics* **174,** 1431–1439.

Zeng, K., Shi, S., and Wu, C-I. (2007) Compound tests for the detection of hitchhiking under positive selection. *Molecular Biology and Evolution* **24(8),** 1898–1908.

Zhao, F. and Qin, W. (2007) Comparative molecular population genetics of phycoerythrin locus in *Prochlorococcus*. *Genetica* **129,** 291–299.