4

Synonymous and Nonsynonymous Nucleotide Substitutions

In chapter 3, we have seen that the rate of nucleotide substitution is much higher at the third positions of codons than at the first and second positions. This is caused by the fact that many nucleotide substitutions at the third positions are silent and do not change amino acids. However, not all substitutions at the third positions are silent. Furthermore, some silent substitutions may also occur at the first positions. It is therefore interesting to know the rates of **synonymous** and **nonsynonymous substitution** separately. Since synonymous substitutions are apparently free from natural selection, the rate of synonymous substitution is often equated to the rate of neutral nucleotide substitution (Miyata et al. 1980). Indeed, the rate of synonymous substitution is similar for many genes, unless it is disturbed by codon usage bias and other factors. By contrast, the rate of nonsynonymous substitution is generally much lower than that of synonymous substitution and varies extensively from gene to gene. This is considered to be due to purifying selection, the extent of which varies from gene to gene (Kimura 1983).

However, it is important to note that there are genes in which nonsynonymous substitutions occur at a higher rate than synonymous substitutions (e.g., Hughes and Nei 1988; Lee et al. 1995). These nonsynonymous substitutions are apparently caused by positive Darwinian selection, because under neutral evolution one would expect that the rates of synonymous and nonsynonymous substitution are equal to each other. For these reasons, estimation of the rates of synonymous and nonsynonymous substitution has become an important subject in the study of molecular evolution.

Estimation of the rates of synonymous and nonsynonymous substitution is more complicated than that of the total number of nucleotide substitutions. In most nucleotide sequences there are more nucleotide sites that potentially produce nonsynonymous mutations than sites that potentially produce synonymous mutations, and the numbers of synonymous and nonsynonymous sites vary from gene to gene. Therefore, the rates of synonymous and nonsynonymous substitution should be defined as the number of synonymous substitutions per synonymous site (r_S) and the number of nonsynonymous substitutions per nonsynonymous site (r_N) per year or per generation. In practice, we usually do not

know the time of divergence (t) between two DNA sequences compared. Therefore, it is customary to consider the number of synonymous substitutions per synonymous site ($d_S = 2r_S t$) and the number of nonsynonymous substitutions per nonsynonymous site ($d_N = 2r_N t$) for a pair of sequences.

There are several methods for estimating d_S and d_N. They can be classified into three groups: (1) evolutionary pathway methods, (2) methods based on Kimura's 2-parameter model, and (3) maximum likelihood methods with codon substitution models. These methods are based on different assumptions, and therefore they do not necessarily give the same results. In this chapter, we explain the first two groups of methods in detail, since they are commonly used in the literature. In the following, we consider the standard genetic code, but the same formulation can be made for any genetic code discussed in chapter 1.

4.1. Evolutionary Pathway Methods

This approach was first used by Miyata and Yasunaga (1980). They considered all possible evolutionary pathways between each pair of homologous codons of two DNA sequences and developed a method for estimating d_S and d_N. However, their method is quite complicated, because every nucleotide substitution is weighted by the likelihood of occurrence of the substitution, taking into account the similarity of amino acids encoded. Conducting a computer simulation, Nei and Gojobori (1986) showed that Miyata and Yasunaga's weighting for different pathways is not necessary and that a simple unweighted version gives essentially the same results as those given by Miyata and Yasunaga. We therefore present Nei and Gojobori's (1986) unweighted pathway method and its modifications.

Nei-Gojobori Method

In Nei and Gojobori's method, d_S and d_N are estimated by computing the numbers of synonymous and nonsynonymous substitutions and the numbers of potentially synonymous and potentially nonsynonymous sites. Let us first consider the numbers of potentially synonymous and potentially nonsynonymous sites. In Nei and Gojobori's method, these numbers are computed for each codon under the assumption of equal probabilities of all nucleotide changes. We denote by f_i the proportion of synonymous changes (the ratio of the number of synonymous changes to the sum of synonymous and nonsynonymous changes, excluding nonsense mutations) at the i-th nucleotide position of a codon ($i = 1, 2, 3$). The numbers of potentially synonymous (s) and potentially nonsynonymous (n) sites for this codon are then given by $s = \Sigma_{i=1}^{3} f_i$ and $n = 3 - s$, respectively. For example, in the case of phenylalanine codon TTT, s becomes

$$s = 0 + 0 + \frac{1}{3} \tag{4.1}$$

because no nucleotide changes at the first and second positions result in synonymous codons and at the third position one out of the three possi-

ble changes results in a synonymous codon (TTC). Since all other changes are nonsynonymous, n is given by $3 - 1/3 = 8/3$. When any nucleotide change results in a termination codon, this change is disregarded. For example, a nucleotide change at the third position of the cysteine codon TGT results in a termination codon when T changes to A, but it gives a synonymous codon when T changes to C and a nonsynonymous codon (Trp) when T changes to G. Therefore, $f_3 = 1/2$ in this case. Since $f_1 = f_2 = 0$ for this codon, we have $s = 0.5$ and $n = 2.5$.

To obtain the total numbers of synonymous sites (S) and nonsynonymous sites (N) for the entire sequence, we use the formulas $S = \Sigma_{j=1}^{C} s_j$ and $N = 3C - S$, where s_j is the value of s for the j-th codon and C is the total number of codons. In practice, we compare two sequences, so the average values of S and N for the two sequences are used in actual computation. Note that $S + N = 3C$ is equal to the total number of nucleotides compared.

Let us now compute the numbers of synonymous and nonsynonymous nucleotide differences between a pair of homologous sequences. We compare the two sequences, codon by codon, and count the number of nucleotide differences for each pair of codons compared. When there is only one nucleotide difference, we can immediately decide whether the difference is synonymous or nonsynonymous. For example, if the codon pairs compared are GTT (Val) and GTA (Val), there is one synonymous difference. We denote by s_d and n_d the numbers of synonymous and nonsynonymous differences per codon, respectively. In the present example, $s_d = 1$ and $n_d = 0$. When two nucleotide differences exist between the two codons compared, there are two possible ways to obtain the differences. For example, in the comparison of TTT and GTA there are two possible parsimonious pathways between the two codons. That is,

(1) TTT (Phe) ↔ GTT (Val) ↔ GTA (Val)
(2) TTT (Phe) ↔ TTA (Leu) ↔ GTA (Val)

Pathway (1) involves one synonymous and one nonsynonymous substitutions, whereas pathway (2) involves two nonsynonymous substitutions. We assume that pathways (1) and (2) occur with equal probability. The numbers of synonymous and nonsynonymous differences then become $s_d = 0.5$ and $n_d = 1.5$, respectively. In some comparisons of codons, there are pathways in which termination codons are involved. We eliminate these pathways from the computation.

When there are three nucleotide differences between the codons compared, there are six different possible pathways between the codons, and in each pathway there are three mutational steps. Considering all these pathways and mutational steps, one can again count the numbers of synonymous and nonsynonymous differences in the same way as in the case of two nucleotide differences. For example, if the two codons compared are TTG and AGA, there are following six pathways.

(1) TTG (Leu) ↔ ATG (Met) ↔ AGG (Arg) ↔ AGA (Arg)
(2) TTG (Leu) ↔ ATG (Met) ↔ ATA (Ile) ↔ AGA (Arg)
(3) TTG (Leu) ↔ TGG (Trp) ↔ AGG (Arg) ↔ AGA (Arg)

(4) TTG (Leu) ↔ TGG (Trp) ↔ TGA (Ter) ↔ AGA (Arg)
(5) TTG (Leu) ↔ TTA (Leu) ↔ ATA (Ile) ↔ AGA (Arg)
(6) TTG (Leu) ↔ TTA (Leu) ↔ TGA (Ter) ↔ AGA (Arg)

Pathways (4) and (6) involve a termination codon, so they are disregarded. The numbers of synonymous substitutions in pathways (1), (2), (3), and (5) are 1, 0, 1, and 1, respectively, whereas the numbers of nonsynonymous substitutions are 2, 3, 2, and 2, respectively. Since we assume that the four pathways are equally probable, we have $s_d = 3/4$ and $n_d = 9/4$.

The total numbers of synonymous and nonsynonymous differences for a sequence comparison can be obtained by summing up these values over all codons. That is, $S_d = \Sigma_{j=1}^{C} s_{dj}$ and $N_d = \Sigma_{j=1}^{C} n_{dj}$, where s_{dj} and n_{dj} are the numbers of synonymous and nonsynonymous differences for the j-th codon, and C is the number of codons compared. Note that $S_d + N_d$ is equal to the total number of nucleotide differences between the two DNA sequences compared.

We can, therefore, estimate the proportions of synonymous (p_S) and nonsynonymous (p_N) differences by the equations

$$\hat{p}_S = S_d/S, \qquad \hat{p}_N = N_d/N \tag{4.2}$$

where S and N are the average numbers of synonymous and nonsynonymous sites for the two sequences compared. To estimate the numbers of synonymous ($\hat{d}_S$) and nonsynonymous ($\hat{d}_N$) substitutions per site, we use the Jukes-Cantor method (Equation [3.8]) replacing p by $\hat{p}_S$ or $\hat{p}_N$. This method, of course, gives only approximate estimates of d_S and d_N, because the nucleotide substitution at the synonymous and nonsynonymous sites does not really follow the Jukes-Cantor model, as noted by Perler et al. (1980). Despite this theoretical problem, computer simulation has shown that Equation (3.8) gives good estimates of synonymous and nonsynonymous substitutions, as long as the nucleotide frequencies of A, T, C, and G are nearly equal and there is no significant transition/transversion bias (Ota and Nei 1994c).

The approximate large-sample variances of $\hat{d}_S$ and $\hat{d}_N$ can be computed by Equation (3.9) if we replace $\hat{p}$ in the equation by $\hat{p}_S$ or $\hat{p}_N$ and n by S or N (Nei 1987). Theoretically, more accurate large-sample variances [$V(\hat{d}_S)$ and $V(\hat{d}_N)$] of $\hat{d}_S$ and $\hat{d}_N$ are given by

$$V(\hat{d}_S) = V(\hat{p}_S)\Big/\left(1 - \frac{4}{3}p_S\right)^2, \qquad V(\hat{d}_N) = V(\hat{p}_N)\Big/\left(1 - \frac{4}{3}p_N\right)^2 \tag{4.3}$$

where

$$V(\hat{p}_S) = \sum_{i=1}^{C}(s_{di} - p_S s_i)^2/S^2, \qquad V(\hat{p}_N) = \sum_{i=1}^{C}(n_{di} - p_N n_i)^2/N^2 \tag{4.4}$$

(Ota and Nei 1994c). However, computer simulation has shown that the above formulas give nearly the same results as those obtained by Equations (3.9).

Another way of computing the variances of $\hat{d}_S$, $\hat{d}_N$, $\hat{p}_S$, and $\hat{p}_N$ is to use

the bootstrap method explained below. As long as S_d, S_n, S, and N are sufficiently large, the bootstrap is expected to give more accurate variances than the above analytical formulas, because it does not depend on the assumption that the expectations of s_{di} and n_{di} are given by $p_S s_i$ and $p_N n_i$, respectively.

Large-Sample Test of the Difference Between $\hat{d}_S$ and $\hat{d}_N$ or Between $\hat{p}_S$ and $\hat{p}_N$

To detect positive Darwinian selection, it is necessary to show that $\hat{d}_N$ is significantly greater than $\hat{d}_S$. A simple way to test the null hypothesis of $d_N = d_S$ is to compute the difference $\hat{D} = \hat{d}_N - \hat{d}_S$ and the variance $[V(\hat{D})]$ of $\hat{D}$ and conduct the normal deviate or the Z test under the assumption that S_d and N_d are sufficiently large (>10) so that the distribution of $\hat{D}$ approximately follows the normal distribution. In the present case, $V(\hat{D})$ is approximately given by $V(\hat{d}_N) + V(\hat{d}_S)$, because $\hat{d}_N$ and $\hat{d}_S$ are theoretically independent of each other. Therefore, we have

$$Z = \hat{D}/s(\hat{D}) \tag{4.5}$$

where $s(\hat{D}) = [V(\hat{D})]^{1/2}$. Here we are interested in $d_N > d_S$, so that the test will be a one-tail test. The Z values for the significance levels of 5, 1, and 0.1% in this case are 1.64, 1.96, and 2.81, respectively. This test corresponds to the t test with an infinite number of degrees of freedom.

The variance of $\hat{D} = \hat{d}_N - \hat{d}_S$ can also be computed by the bootstrap method. In the bootstrap method, a pair of random codon sequences consisting of the same number of codons as that of the original sequences are generated by the resampling method described in chapter 2. In the present case, however, codons rather than nucleotides are the units of resampling. For the b-th bootstrap sample of codon sequences, we compute estimates ($\hat{p}_{Sb}$, $\hat{p}_{Nb}$, $\hat{d}_{Sb}$, and $\hat{d}_{Nb}$) of p_S, p_N, d_S, and d_N. Therefore, if we repeat this computation about 1000 times, we can compute the variances of these quantities using Equation (2.16). For testing the observed difference $\hat{D} = \hat{d}_N - \hat{d}_S$, we do not really need the variances of $\hat{d}_S$ and $\hat{d}_N$. Instead, we can compute the standard error of $\hat{D}$ directly by using the bootstrap and then use the Z test. An even simpler method would be to compute the number (B_S) of bootstrap replications in which the bootstrap estimate ($\hat{D}_b$) of D is smaller than 0 and compute the proportion of these replications, B_S/B, where B is the total number of bootstrap replications. This proportion is called the **achieved significance level** (*ASL*) to distinguish it from the model-based significance level (Efron and Tibshirani 1993). If *ASL* is less than 5 or 1%, one may conclude that the observed $\hat{D}$ is significantly greater than 0. However, this test seems to be less accurate than the Z test (Efron and Tibshirani 1993).

When the nucleotide sequences are short, $\hat{p}_S$ or $\hat{p}_N$ can be greater than 0.75 by chance or for some other reasons, and $\hat{d}_S$ or $\hat{d}_N$ may not be computable. In this case, the difference between synonymous and nonsynonymous substitutions should be tested by using $\hat{p}_S$ and $\hat{p}_N$ directly. Since the variances of $\hat{p}_S$ and $\hat{p}_N$ can be computed either by Equation (4.4) or by the bootstrap, the null hypothesis of $p_S = p_N$ can be tested by using

Table 4.1 Fisher's exact test for small samples.

	Substitution Sites	Nonsubstitution Sites	Total
Synonymous	S_d (1)	$S - S_d$ (40)	S (41)
Nonsynonymous	N_d (20)	$N - N_d$ (110)	N (130)
Sum	$S_d + N_d$ (21)	$T - S_d - N_d$ (150)	T (171)

Note: The numbers in parentheses refer to those used for testing adaptive evolution at a human MHC (*HLA-A*) locus. $T = S + N$.

the Z test. Actually, $\hat{p}_S$ and $\hat{p}_N$ are better than $\hat{d}_S$ and $\hat{d}_N$ in detecting positive selection, because they require fewer assumptions than the latter.

Small-Sample Test

When the number of nucleotide substitutions per sequence (S_d or N_d) is small, the above large-sample test tends to be too liberal and may be misleading (Zhang et al. 1997). In this case, it is usually possible to count the actual numbers of synonymous (S_d) and nonsynonymous (N_d) substitutions without much error, because most codon differences are caused by one nucleotide substitution. We can then construct a 2×2 contingency table for synonymous and nonsynonymous substitutions and conduct Fisher's exact test as given in Table 4.1. In this table, T stands for the total number of nucleotides examined, i.e., $T = S + N$. An example of Fisher's exact test will be discussed later.

Tests of the Difference Between $\bar{d}_S$ and $\bar{d}_N$ or $\bar{p}_S$ and $\bar{p}_N$

In the study of adaptive evolution, it is often necessary to compare the average values ($\bar{d}_S$ and $\bar{d}_N$) of $\hat{d}_S$'s and $\hat{d}_N$'s for many sequence comparisons, because comparison of a pair of sequences is not always very informative. Hughes and Nei (1988, 1989) used this approach to show that $\bar{d}_N$ is significantly greater than $\bar{d}_S$ in the antigen recognition site of major histocompatibility complex (MHC) genes, and this demonstration led them to conclude that the extremely high degree of polymorphism at MHC loci is primarily due to overdominant selection operating at the antigen recognition site.

To establish $\bar{d}_N > \bar{d}_S$, however, it is necessary to conduct a statistical test of the difference $\hat{D} = \bar{d}_N - \bar{d}_S$. This test can be done by using the standard Z test with the variance $[V(\hat{D})]$ of $\hat{D}$ given by $V(\bar{d}_N) + V(\bar{d}_S) - 2Cov(\bar{d}_N, \bar{d}_S)$, where $V(\bar{d}_N)$, $V(\bar{d}_S)$, and $Cov(\bar{d}_N, \bar{d}_S)$ are the variances and covariance of $\bar{d}_N$ and $\bar{d}_S$. It is not a simple matter to compute $V(\bar{d}_N)$, $V(\bar{d}_S)$, and $Cov(\bar{d}_N, \bar{d}_S)$ analytically, because different sequences are related through evolutionary history. Nei and Jin (1989) developed a method for computing the variances and covariance of $\bar{d}_N$ and $\bar{d}_S$ taking into account the phylogenetic tree of the sequences. This method is useful when the number of sequences used is relatively small but becomes time consuming when the number is large (say, more than 20). Another method for testing the difference $\hat{D}$ is to use the bootstrap method.

Suppose that all sequences consist of C codons. We can then resample C codons with replacement from the original set of sequences and compute $\bar{d}_N$ and $\bar{d}_S$ for this set of samples. If we repeat this computation many times, we can compute the standard error of $\bar{d}_N - \bar{d}_S$ and use it for the Z test. If one is interested in the test of the mean difference $\bar{p}_N - \bar{p}_S$, the same method can be used.

It should be noted that the bootstrap test may lead to an erroneous conclusion when the numbers of synonymous and nonsynonymous substitutions observed are small. To explain this problem, let us consider an extreme case where 6 nonsynonymous ($n = 6$) and 0 synonymous substitutions ($s = 0$) are observed when 60 nonsynonymous sites ($N = 60$) and 30 synonymous sites ($S = 30$) are examined. In this case, Fisher's exact test mentioned above indicates that the null hypothesis $p_N = p_S$ cannot be rejected. However, if we use the bootstrap test, $\hat{p}_N$ would be greater than $\hat{p}_S$ in almost all replications. Therefore, we would conclude that the null hypothesis is rejected. This obviously incorrect conclusion was reached because the original values of s and n were biased by chance and this bias cannot be corrected by bootstrap resampling. It is therefore important to compute the standard error of $\hat{D}$ by analytical formulas when C is small.

Modified Nei-Gojobori Method

Nei and Gojobori's (1986) method assumes random nucleotide substitution among the four nucleotides in computing the number of synonymous and nonsynonymous sites. In practice, this assumption does not necessarily hold, and the rate of transitional change is usually higher than that of transversional change. In this case, the number (S) of potential sites that can produce synonymous substitutions is expected to be greater than the number estimated by Nei and Gojobori's method, because transitional changes at third positions are largely synonymous. Therefore, Nei and Gojobori's method is expected to give overestimates of p_S and d_S and underestimates of p_N and d_N.

To rectify this deficiency, Ina (1995) proposed a method for estimating d_S and d_N using Kimura's (1980) 2-parameter model. His method is quite elaborate, as will be explained later. However, the major problem of Nei and Gojobori's method is the underestimation of S and the overestimation of N. Therefore, if we use appropriate methods of estimating S and N, their approach can still be used (Zhang et al. 1998). In the following, we adapt Ina's method for this purpose.

In Kimura's (1980) model the rates of transitional and transversional changes are given by α and β, respectively (chapter 3), but since any nucleotide can have two different transversional changes, the proportion of transitions among the total changes is given by

$$\frac{\alpha}{\alpha + 2\beta} = \frac{R}{1 + R} \tag{4.6}$$

where R is the transition/transversion ratio and becomes 0.5 when there is no bias. (Note that R is different from the transition/transversion rate

ratio $k = \alpha/\beta$, which is often used in theoretical papers.) Ina (1995) has shown that the expected number of synonymous changes per codon can be expressed in terms of $R = \alpha/(2\beta)$ for all codons. For example, for codon TTT the number is given by

$$s = 0 + 0 + \frac{\alpha}{\alpha + 2\beta} = \frac{R}{1 + R} \tag{4.7}$$

because in this case only the third nucleotide position produces synonymous changes and only one (T → C) of the three possible changes is synonymous. For another example, codon CTA (Leu) has the expected number of $s = R/(1 + R) + 1$, because the first, second, and third nucleotide positions of this codon can produce synonymous substitutions with probabilities $R/(1 + R)$, 0, and 1, respectively. In these computations, nonsense mutations are disregarded as before.

It is therefore clear that if we know R, we can compute s for all codons and then estimate S and N $(= 3C - S)$. The problem is how to estimate R from actual data. We suggest that R be estimated by Equation (3.2) or (3.18) in chapter 3 or that the R value obtained from other information be used. Theoretically, when the pattern of nucleotide substitution is complicated, both Equations (3.2) and (3.18) may give underestimates of R (Yang 1995b), and this underestimation of R makes the test of positive selection conservative. However, it is better to use a conservative test for detecting positive selection, because the actual pattern of nucleotide substitution is usually quite complicated and this may inflate $\hat{d}_N$ relative to $\hat{d}_S$ spuriously.

If we use the above method, S is expected to increase, and N is expected to decrease compared with the values obtained by the original Nei-Gojobori method. Let us denote these new S and N by S_R and N_R, respectively. In contrast to S and N, the number of synonymous (S_d) and nonsynonymous (N_d) differences are not seriously affected by the transition/transversion bias, because S_d and N_d are based on the actual number of substitutions observed. Therefore, the proportions of synonymous ($\hat{p}_S$) and nonsynonymous ($\hat{p}_N$) differences are now given by

$$\hat{p}_S = S_d/S_R, \qquad \hat{p}_N = N_d/N_R \tag{4.8}$$

whereas the estimates ($\hat{d}_S$ and $\hat{d}_N$) of d_S and d_N are again approximately given by the Jukes-Cantor formula. Theoretically, there is a better way to estimate d_S and d_N as shown by Ina (1995), but in practice there is not much difference between the estimates obtained by the two methods unless d_S and d_N are very high. (When $d_S > 1.0$ and $d_N > 1.0$, the reliability of $\hat{d}_S$ and $\hat{d}_N$ is very low, because the actual process of synonymous and nonsynonymous substitution is very complicated.) Furthermore, the present methods give smaller variances of $\hat{p}_S$, $\hat{p}_N$, $\hat{d}_S$, and $\hat{d}_N$ than those obtained by Ina's method.

Although the modified Nei-Gojobori method is theoretically better than the original version when Kimura's model with a high R value applies, it should be noted that when the estimate of R is unreliable, it may lead to an erroneous conclusion. Particularly when an overestimate of R

is used, the modified version may conclude that $\hat{d}_N$ is significantly higher than $\hat{d}_S$, even if this is not actually the case. Note that the actual pattern of nucleotide substitution is much more complicated than the Kimura model, and under certain conditions the modified Nei-Gojobori method may give an overestimate of S and an underestimate of N. For this reason, it is always better to use both the original Nei-Gojobori method and the modified version to detect positive selection. If the original version indicates positive Darwinian selection, the conclusion would be safer.

Example 4.1. Positive Darwinian Selection at MHC Loci

Figure 4.1 shows the nucleotide sequences of three alleles from the A locus of the human MHC (HLA) class I α chain genes. The α chain gene encodes three extracellular domains (α1, α2, and α3), a transmembrane portion, and a cytoplasmic tail of the MHC molecule (Klein and Horejsi 1997), and the sequences in Figure 4.1 are for the α1, α2, and α3 extracellular domains. They consist of $C = 274$ codons or $3C = 822$ nucleotides. Comparison of alleles *A*2301* and *A*2501* shows that there are 41 nucleotide differences; 33 of them are from codons showing one nucleotide difference and eight are from codons showing two nucleotide differences. All the codon differences and the s and n values for each codon difference are presented in Table 4.2. From this table, we obtain $S_d = 11.5$ and $N_d = 29.5$.

Nei-Gojobori Method

The total number of synonymous sites (S) can be computed by the methods described above, and it becomes 198 and 195.8 for alleles *A*2301* and *A*2501*, respectively. Therefore, the average of S for the two sequences is 196.9, and the average N is $822 - 196.9 = 625.1$. We can then obtain $\hat{p}_S = 11.5/196.9 = 0.0584$ and $\hat{p}_N = 29.5/625.1 = 0.0472$ from Equation (4.2), and their standard errors become $s(\hat{p}_S) \equiv [V(\hat{p}_S)]^{1/2} = 0.0167 = s(\hat{p}_N) = 0.0085$ from Equation (4.4). Essentially the same standard errors (0.0160 and 0.0087, respectively) are obtained by the bootstrap method. The Z value equivalent to Equation (4.5) is -0.60, which indicates that the difference $\hat{p}_N - \hat{p}_S$ ($= -0.011$) is not statistically significant. (Here the two-tail test should be used.) If we use Equations (3.8) and (4.3), we obtain $\hat{d}_S = 0.0608 \pm 0.0181$ and $\hat{d}_N = 0.0487 \pm 0.0091$ (Table 4.3). The Z value for the difference $\hat{d}_N - \hat{d}_S$ then becomes -0.60, which again indicates that the difference is not significant. Therefore, the Z tests for $\hat{d}_N - \hat{d}_S$ and $\hat{p}_N - \hat{p}_S$ give the same conclusion.

Modified Nei-Gojobori Method

In this method, we first have to estimate the R value. If we use Equation (3.18), we have $R = 0.79$ for alleles *A*2301* and *A*2501*, $R = 0.92$ for alleles *A*2301* and *A*3301*, and $R = 0.82$ for alleles *A*2501* and *A*3301*. Therefore, the average R is approximately 0.85. Using this R value, we have S_R equal to 211.9 and 209.8 for alleles *A*2301* and *A*2501*, respectively, with an average of 210.8. This gives $N_R = 611.2$. Using these

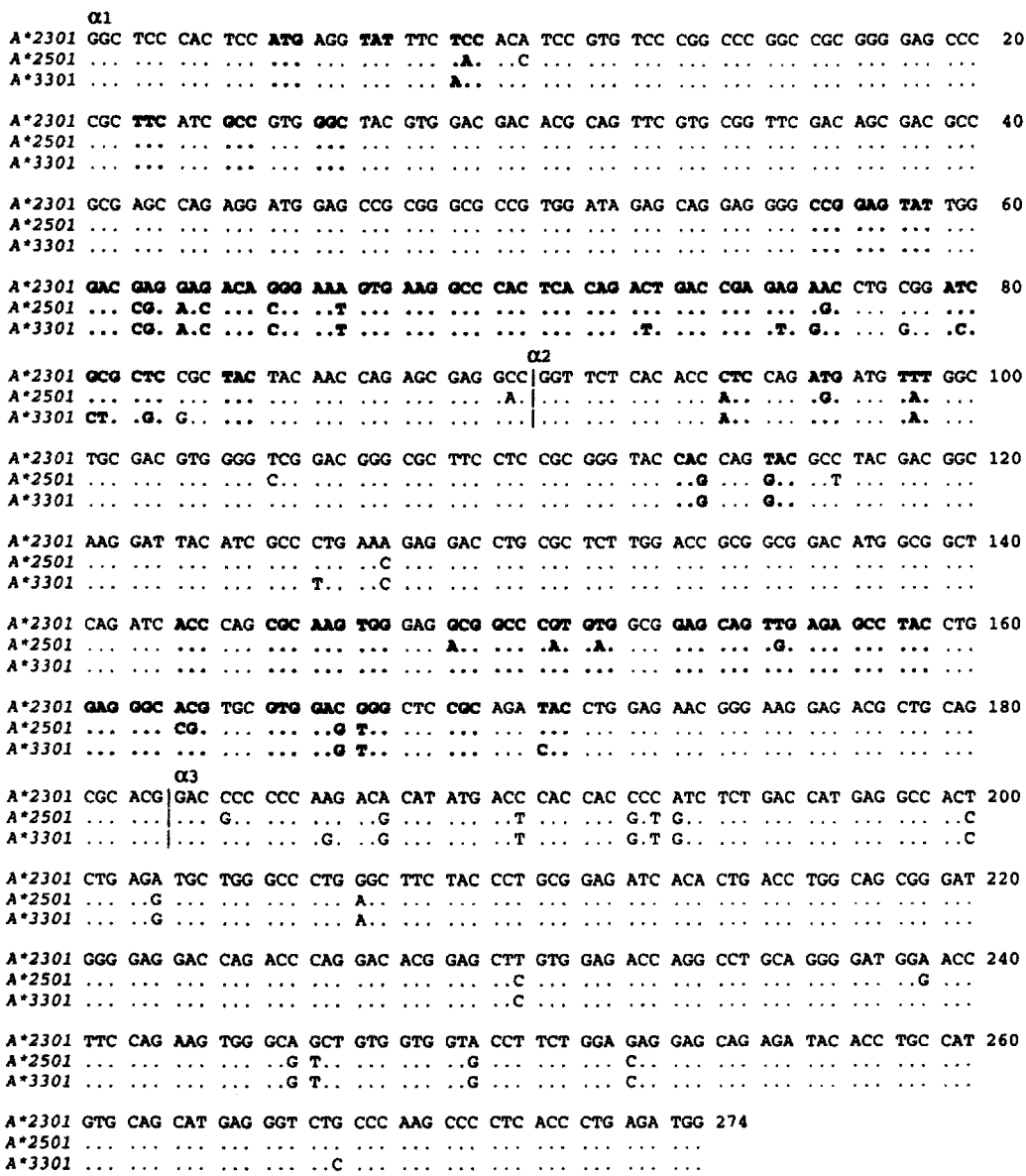

FIGURE 4.1. Nucleotide sequences of three human class I *HLA-A* alleles for the three extracellular domains α_1, α_2, and α_3. A dot (.) shows identity with the first sequence. Exons boundaries are marked with vertical lines. The nucleotides at the antigen recognition site (ARS) are in boldface.

values, we obtain $\hat{d}_S = 0.0566 \pm 0.0169$ and $\hat{d}_N = 0.0499 \pm 0.0093$. Therefore, $\hat{d}_S$ has decreased and $\hat{d}_N$ has increased slightly.

Adaptive Evolution

X-ray diffraction studies have shown that class I MHC molecules form a groove in which a foreign peptide is bound (Bjorkman et al. 1987a,

Table 4.2 Codons that are different between the HLA *A*2301* and *A*2501* alleles.

Codon		s_d	n_d	$s_d + n_d$	Codon		s_d	n_d	$s_d + n_d$
*9	TCC-TAC		1	1	*156	TTG-TGG		1	1
10	ACA-ACC	1		1	*163	ACG-CGG	0.5	1.5	2
*62	GAG-CGG		2	2	*166	GAC-GAG		1	1
*63	GAG-AAC		2	2	*167	GGG-TGG		1	1
*65	GGG-CGG		1	1	184	CCC-GCC		1	1
*66	AAA-AAT		1	1	187	ACA-ACG	1		1
*77	AAC-AGC		1	1	190	ACC-ACT	1		1
90	GCC-GAC		1	1	193	CCC-GCT	1	1	2
*95	CTC-ATC		1	1	194	ATC-GTC		1	1
*97	ATG-AGG		1	1	200	ACT-ACC	1		1
*99	TTT-TAT		1	1	202	AGA-AGG	1		1
105	TCG-CCG		1	1	207	GGC-AGC		1	1
*114	CAC-CAG		1	1	230	CTT-CTC	1		1
*116	TAC-GAC		1	1	239	GGA-GGG	1		1
117	GCC-GCT	1		1	245	GCA-GCG	1		1
127	AAA-AAC		1	1	246	GCT-TCT		1	1
*149	GCG-ACG		1	1	249	GTA-GTG	1		1
*151	CGT-CAT		1	1	253	GAG-CAG		1	1
*152	GTG-GAG		1	1		Total	11.5	29.5	41

Note: Antigen recognition sites are indicated with an asterisk.

1987b). This groove is called the antigen recognition site (ARS) and consists of 57 amino acid sites (boldfaced letters in Figure 4.1). If we apply the Nei-Gojobori method to the 57 amino acid sites of the ARS for alleles *A*2301* and *A*2501*, we obtain $S_d = 0.5$, $N_d = 20.5$, $S = 40.5$, and N

Table 4.3 Numbers of synonymous ($\hat{d}_S$) and nonsynonymous ($\hat{d}_N$) substitutions between the HLA *A*2301* and *A*2501* alleles for the extracellular region and the antigen recognition sites (ARS).

	Extracellular Region ($C = 274$)		ARS ($C = 57$)	
Method	$\hat{d}_S$	$\hat{d}_N$	$\hat{d}_S$	$\hat{d}_N$
$R = 0.5$				
NG[a]	6.08 ± 1.81	4.87 ± 0.91	1.24 ± 1.76	17.63 ± 4.03
LWL[b]	6.52 ± 2.02	4.82 ± 0.89	0.03 ± 1.87	17.25 ± 3.99
$R = 0.85$[c]				
Modified-NG	5.66 ± 1.69	4.99 ± 0.93	1.17 ± 1.66	18.06 ± 4.14
PBL[d]	4.59 ± 1.46	4.80 ± 0.90	0.02 ± 1.14	17.04 ± 3.96
Kumar	4.55 ± 1.46	4.74 ± 0.91	0.36 ± 1.19	16.79 ± 4.03
Ina II	4.87 ± 1.47	5.31 ± 0.99	1.50 ± 2.13	16.67 ± 3.81
GY[e]	12.17	4.25	0.02	16.98

Note: $\hat{d}_S$ and $\hat{d}_N$ are multiplied by 100.

[a]NG: Nei-Gojobori.
[b]LWL: Li-Wu-Luo.
[c]$R = 0.85$ was used only for the Modified-NG method. In the other methods, R was computed automatically.
[d]PBL: Pamilo-Bianchi-Li.
[e]GY: Goldman-Yang.

= 130.5. We therefore have $\hat{d}_S = 0.0124 \pm 0.0176$ and $\hat{d}_N = 0.1763 \pm 0.0403$ (Table 4.3). A Z test shows that $Z = 3.7$ and $\hat{d}_N$ is significantly greater than $\hat{d}_S$ at the 0.1% level when a one-tail test is used. In contrast, the modified Nei-Gojobori method gives $S_R = 43.11$ and $N_R = 127.89$ (with $R = 0.85$), so that we have $\hat{d}_S = 0.0117 \pm 0.0166$ and $\hat{d}_N = 0.1806 \pm 0.0414$. The Z test again shows that $\hat{d}_N$ is significantly greater than $\hat{d}_S$ at the 0.1% level. Therefore, both methods show that $\hat{d}_N$ is greater than $\hat{d}_S$, and this strongly suggests that the ARS of class I MHC molecules is the target of positive Darwinian selection.

Small-Sample Tests

In the above tests, we used a large-sample test, which is not really valid because S_d was only 0.5. Let us now use Fisher's exact test. If we use the conservative Nei-Gojobori method, we obtain $S = 41$ and $N = 130$ approximately. We also assume $S_d = 1$ and $N_d = 20$ to make the test even more conservative. We then have the 2×2 contingency table given in parentheses in Table 4.1. Fisher's exact test gives a P value of 0.018. This indicates that $\hat{d}_N$ is significantly greater than $\hat{d}_S$. If we use the modified Nei-Gojobori method, we obtain $S_R = 44$ and $N_R = 127$ (with $R = 0.85$), and Fisher's test gives a P value of 0.012. Therefore, the P values for the small-sample test are higher than those for the large-sample test.

Tests of $\bar{d}_N - \bar{d}_S$ or $\bar{p}_N - \bar{p}_S$

Since the power of detecting positive selection is low in this case because of the small number of codons involved, let us consider the averages of $\hat{d}_N$ and $\hat{d}_S$. In the present case, there are three sequences, so $\hat{d}_N$ and $\hat{d}_S$ can be computed for three pairs of alleles. We can then obtain the averages ($\bar{d}_N$ and $\bar{d}_S$) of these values. If we use the Nei-Gojobori method, they become $\bar{d}_N = (0.1763 + 0.1822 + 0.1479)/3 = 0.1688$ and $\bar{d}_S = (0.0124 + 0.0000 + 0.0124)/3 = 0.0083$. Therefore, the difference $\hat{D} = \bar{d}_N - \bar{d}_S$ is 0.1605. If we use the bootstrap method, the standard error of $\hat{D} = \bar{d}_N - \bar{d}_S$ becomes 0.0322. (This was computed by a program in MEGA2 using 1000 bootstrap replications.) Therefore, we have $Z = 4.98$. If we use Nei and Jin's method, we obtain $Z = 4.80$, which is again highly significant. These results reinforce the conclusion reached by comparison of two sequences.

4.2. Methods Based on Kimura's 2-Parameter Model

Li-Wu-Luo Method

Li et al. (1985) developed another method, based on Kimura's 2-parameter model. They first noted that when the degeneracy of the genetic code is considered, the nucleotide sites of codons can be classified into 4-fold degenerate, 2-fold degenerate, and 0-fold degenerate (nondegenerate) sites with a few exceptions (e.g., isoleucine codons). A site is called 4-

fold degenerate if all possible changes at the site are synonymous, 2-fold degenerate if one of the three possible changes is synonymous, and 0-fold degenerate if all changes are nonsynonymous or nonsense mutations. For example, the third nucleotide positions of the valine codons are 4-fold degenerate sites, and the second positions of all codons are 0-fold degenerate sites. The third positions of the three isoleucine codons are actually 3-fold degenerate sites, but they are regarded as 2-fold degenerate sites to simplify the computation.

Using the above rule, we can compute the numbers of three types of sites for each of the two sequences and denote by L_0, L_2, and L_4 the average numbers of 0-fold, 2-fold, and 4-fold degenerate sites for the two sequences compared, respectively. We then compare the two sequences, codon by codon, and classify each nucleotide difference as either a transition or a transversion. We denote by P_i and Q_i the proportions of transitional and transversional nucleotide differences at the i-th class of nucleotide sites (i = 0, 2, or 4). (Actually, they considered all possible evolutionary pathways between each pair of codons as in the case of the Nei-Gojobori method and computed P_i and Q_i taking into account the likelihood of occurrence of each amino acid substitution. See Li et al. [1985] for the detail.) We can then estimate the numbers of transitional (A_i) and transversional (B_i) substitutions per site for each of the three classes of nucleotide sites. That is,

$$A_i = \frac{1}{2}\ln(a_i) - \frac{1}{4}\ln(b_i) \tag{4.9a}$$

$$B_i = \frac{1}{2}\ln(b_i) \tag{4.9b}$$

where $a_i = 1/(1 - 2P_i - Q_i)$ and $b_i = 1/(1 - 2Q_i)$.

We note that all substitutions at 4-fold sites are synonymous and all substitutions at 0-fold sites are nonsynonymous. At 2-fold sites, transitional changes (A_2) are mostly synonymous, whereas transversional changes are mostly nonsynonymous. Assuming that nucleotide substitution occurs with equal frequency among the four nucleotides A, T, C, and G, Li et al. (1985) suggested that one third of 2-fold degenerate sites are potentially synonymous sites and two thirds are potentially nonsynonymous sites. With this assumption, they proposed that d_S and d_N be estimated by the following formulas.

$$\hat{d}_S = \frac{3[L_2A_2 + L_4(A_4 + B_4)]}{L_2 + 3L_4} \tag{4.10a}$$

$$\hat{d}_N = \frac{3[L_0(A_0 + B_0) + L_2B_2]}{3L_0 + 2L_2} \tag{4.10b}$$

These formulas depend on a number of assumptions, which are not always satisfied with actual data. First, the type of a given nucleotide site in one sequence may not be the same as that of the homologous site in the other sequence. For example, the type of a given position in one se-

quence may be 2-fold degenerate, but the type of the same position in the other sequence could be 4-fold degenerate. This can happen quite often when sequence divergence is high. In this case, one half of the site is regarded as a 2-fold degenerate site, and the other half as a 4-fold degenerate site. Second, nonsense mutations are counted as nonsynonymous changes. For example, a nucleotide substitution at the third position of tyrosine codon TAT may produce one synonymous codon (TAC) and two nonsense codons (TAA and TAG), but the latter two changes are regarded as nonsynonymous changes. Since nonsense mutations occur with a probability of nearly 4% (chapter 1), this method is expected to give overestimates of d_N. Third, the transitions at the first nucleotide position of four 2-fold degenerate arginine codons (CGA, CGG, AGA, and AGG) are not synonymous but all nonsynonymous with one exception (CGA) that results in a nonsense codon. At the third position of the three isoleucine codons that are 3-fold degenerate, some transversions are synonymous. Despite these problems, Li et al.'s (1985) method seems to give results similar to those obtained by the Nei-Gojobori method when the number of codons is large and sequence divergence is low. When the number of codons used is small (say <100), however, Li et al.'s method may give negative estimates, because a_i and b_i in Equation (4.9) are subject to large sampling errors.

Pamilo-Bianchi-Li Method

Another problem in Li et al.'s (1985) method is the effect of transition/transversion bias, and the error introduced by this bias may be substantial when R is high, as in the case of Nei and Gojobori's method. For this reason, Pamilo and Bianchi (1993) and Li (1993) independently extended Li et al.'s method.

Noting that synonymous transitional changes occur only at 2-fold and 4-fold sites in Li et al.'s model, they proposed that the total number of these changes be estimated by the weighted mean $(L_2A_2 + L_4A_4)/(L_2 + L_4)$. Since the transversions at 4-fold sites are also synonymous, the total number of synonymous substitutions per synonymous site is now estimated by

$$\hat{d}_S = (L_2A_2 + L_4A_4)/(L_2 + L_4) + B_4 \qquad (4.11a)$$

Using the same argument, they also suggested that d_N be estimated by

$$\hat{d}_N = A_0 + (L_0B_0 + L_2B_2)/(L_0 + L_2) \qquad (4.11b)$$

Comeron and Kumar Methods

As mentioned earlier, the treatment of arginine and isoleucine codons in Li et al.'s (1985) method is inaccurate. This is also true with the Pamilo-Bianchi-Li method. This creates a problem when these amino acids are abundant. (In mammalian protamine P1, about 50% of amino acids are arginines; Rooney et al. 2000 n.d.). Comeron (1995) attempted to solve this problem by dividing 2-fold degenerate sites into two groups: 2S-fold

and 2V-fold degenerate sites. The former refer to sites where the two transitional changes are synonymous and the transversional change is nonsynonymous, whereas the latter represent sites where the transitional change is nonsynonymous and the two transversional changes are synonymous. This subdivision of 2-fold degenerate sites certainly help to correct some of Li et al.'s inaccurate classifications of synonymous and nonsynonymous sites (e.g., methionine codons).

However, this does not solve all the problems. For example, a mutation at the first nucleotide position of arginine codon CGG produces TGG (Trp), AGG (Arg), or GGG (Gly). In this case, the transitional change (C → T) results in a nonsynonymous substitution, whereas one transversional change (C → A) results in a synonymous substitution and the other transversional change (C → G) a nonsynonymous substitution. Therefore, this nucleotide site is neither a 2S-fold nor a 2V-fold site. Similarly, the first position of three arginine codons (CGU, CGC, and CGA) and the third position of two isoleucine codons (ATT and ATC) cannot be assigned to any of Comeron's categories.

To take care of these problems, S. Kumar (n.d.) developed another version of the Pamilo-Bianchi-Li method. In this version, nucleotide sites are first classified into 0-fold, 2-fold, and 4-fold degenerate sites, and the 2-fold degenerate sites are further subdivided into simple 2-fold and complex 2-fold degenerate sites. Simple 2-fold degenerated sites are those at which the transitional change results in a synonymous substitution and the two transversional changes generate nonsynonymous substitutions or nonsense mutations. All other 2-fold degenerate sites, including those for the three isoleucine codons, belong to the complex 2-fold sites. Using this classification of sites, Kumar developed a new method for estimating d_S and d_N. This method is included in MEGA2.

Ina's Method

Ina (1995) developed yet another method for estimating d_S and d_N, combining some features of the original Nei-Gojobori and the Pamilo-Bianchi-Li methods. He proposed two methods: method I and method II. In method I, the transition/transversion rate ratio $k = \alpha/\beta$ is estimated by Equation (3.18) in chapter 3 using only data at the third codon positions. This depends on the assumption that the nucleotide substitution at the third positions is largely neutral. S and N are then estimated by using the procedure of the modified Nei-Gojobori method, whereas S_d and N_d are computed by the Nei-Gojobori method. However, Ina divides S_d into synonymous transitional differences (S_{Ts}) and synonymous transversional differences (S_{Tv}) and N_d into nonsynonymous transitional differences (N_{Ts}) and nonsynonymous transversional differences (N_{Tv}). He then estimates $\hat{d}_S$ and $\hat{d}_N$ using formulas analogous to Equation (3.12). In his method II, S and N are estimated from data for all three codon positions, but α and β are estimated by using only synonymous substitutions to reflect the mutation rates before selection. The actual procedure is quite elaborate.

Ina's computer simulation has shown that method II gives slightly more accurate estimates of d_S and d_N than method I when the number of

nucleotides used is large. However, the differences in $\hat{d}_S$ and $\hat{d}_N$ between the two methods or between Ina's methods and the modified Nei-Gojobori method are usually small. Furthermore, when the number of nucleotides used is small and sequence divergence is low, Ina's method I may not be applicable, because no transitions or no transversions may be observed and this makes the estimate of α/β either 0 or ∞. Therefore, some caution is necessary when Ina's methods are to be used.

Example 4.2. Further Analysis of MHC Gene Sequences

In Example 4.1, we computed $\hat{p}_S$, $\hat{p}_N$, $\hat{d}_S$, and $\hat{d}_N$ for human MHC alleles by using the Nei-Gojobori and the modified Nei-Gojobori methods. Let us now compute $\hat{d}_S$ and $\hat{d}_N$ for the alleles *A*2301* and *A*2501* using the methods based on Kimura's model. We will not consider $\hat{p}_S$ and $\hat{p}_N$, since these are not computable in these methods. In the case of the Li-Wu-Luo method, comparison of the two alleles gives $L_0 = 535$, $L_2 = 154.5$, and $L_4 = 132.5$. We also obtain $A_0 = 0.01542$, $B_0 = 0.02985$, $A_2 = 0.00806$, $B_2 = 0.42183$, $A_4 = 0.07359$, and $B_4 = 0.00761$. Therefore, Equations (4.10a) and (4.10b) give $\hat{d}_S = 0.0652 \pm 0.0202$ and $\hat{d}_N = 0.0482 \pm 0.0089$. These values are nearly the same as those obtained by the Nei-Gojobori method (Table 4.3). The Pamilo-Bianchi-Li method gives $\hat{d}_S = 0.0459 \pm 0.0146$ and $\hat{d}_N = 0.0480 \pm 0.0090$. The $\hat{d}_S$ and $\hat{d}_N$ values obtained by the Kumar method and the Ina method II are presented in Table 4.3. The values by the Kumar method are similar to those obtained by the Pamilo-Bianchi-Li method, whereas those by the Ina method II are similar to those obtained by the modified Nei-Gojobori method. Curiously, the $\hat{d}_N$ values obtained by the Pamilo-Bianchi-Li and the Kumar methods are similar to the $\hat{d}_N$ obtained by the Nei-Gojobori method, though they are supposed to be higher than the latter because the transition/transversion bias is taken care of. This is probably caused by the fact that the pattern of nucleotide substitution in MHC genes is much more complicated than Kimura's model (Hughes and Nei 1988).

Let us now compute $\hat{d}_S$ and $\hat{d}_N$ for the ARS using the Li-Wu-Luo, Pamilo-Bianchi-Li, and Kumar methods. The results obtained are presented in Table 4.3 together with the previous results. The $\hat{d}_N$ value obtained by the Li-Wu-Luo method is similar to that obtained by the Nei-Gojobori method, but the $\hat{d}_S$ is much smaller than that obtained by the latter method. This small $\hat{d}_S$ is unreasonable, because the Nei-Gojobori method is based on a parsimonious counting of synonymous substitutions, and therefore it should give a minimum estimate. The unduly low $\hat{d}_S$ value probably occurred because the Kimura model is unlikely to apply to the ARS, where the pattern of nucleotide substitution is complicated and the number of codons involved is small. An unduly small $\hat{d}_S$ value is also obtained by the Pamilo-Bianchi-Li and the Kumar methods, which are also based on the Kimura model. The $\hat{d}_N$ value obtained by the Li-Wu-Luo method is similar to that obtained by the Nei-Gojobori method, but the values obtained by the Pamilo-Bianchi-Li and the Kumar methods are smaller than the $\hat{d}_N$ obtained by the modified Nei-Gojobori method and are virtually the same as the $\hat{d}_N$ obtained by the Li-Wu-Luo method, although they are supposed to be larger. They are even smaller

Table 4.4 Synonymous ($\hat{d}_S$) and nonsynonymous ($\hat{d}_N$) substitutions for the mitochondrial *Nd5* gene sequences from humans and chimpanzees.

Method	$\hat{d}_S$	$\hat{d}_N$
$R = 0.5$		
Nei-Gojobori	41.51 ± 3.80	3.79 ± 0.54
Li-Wu-Luo	42.77 ± 4.14	3.78 ± 0.54
$R = 9.21^a$		
Modified Nei-Gojobori	27.30 ± 2.40	4.38 ± 0.62
Ina II	30.31 ± 3.07	4.38 ± 0.63
Pamilo-Bianchi-Li	30.18 ± 2.87	4.38 ± 0.63
Comeron-Kumar	30.18 ± 2.87	4.38 ± 0.63
Goldman-Yang	28.72	4.42

Note: $\hat{d}_S$ and $\hat{d}_N$ are multiplied by 100.

[a] $R = 9.21$ was used only for the Modified-NG method. In the other methods, R is computed automatically. The number of codons used (C) is 603.

than the $\hat{d}_N$ obtained by the parsimonious Nei-Gojobori method. This again suggests that the Kimura model is inappropriate for MHC genes, particularly for the ARS. Table 4.3 also includes the $\hat{d}_S$ and $\hat{d}_N$ obtained by Ina's method II. These values are more similar to those obtained by the Nei-Gojobori method than those obtained by the modified Nei-Gojobori method. This also suggests that the Kimura model is inappropriate for the ARS.

Example 4.3. $\hat{d}_S$ and $\hat{d}_N$ Values for the Mitochondrial *Nd5* Gene

In the above example, the transition/transversion bias ($R = 0.85$) was small, so that the differences in $\hat{d}_S$ and $\hat{d}_N$ between the two assumptions of $R = 0.5$ and $R = 0.85$ were also small. In many nuclear genes, R is 0.5 ~ 2, and the effect of the transition/transversion bias is usually very small. In mitochondrial genes, however, the effect is expected to be large because R is generally high. Let us now consider the mitochondrial NADH dehydrogenase 5 (*Nd5*) gene sequences from humans and chimpanzees (Horai et al. 1995). The total number of codons in this gene is 603, and Equation (3.18) gives an R value of 9.21. The $\hat{d}_S$ and $\hat{d}_N$ values were obtained by the seven different methods discussed above, and they are presented in Table 4.4. In this case, the assumption of $R = 0.5$ certainly gives overestimates of d_S and underestimates of d_N. However, different methods that allow a high R value give very similar estimates of d_S and d_N.

4.3. Nucleotide Substitutions at Different Codon Positions

When relatively closely related species are compared, the number of synonymous substitutions is expected to increase almost linearly with time,

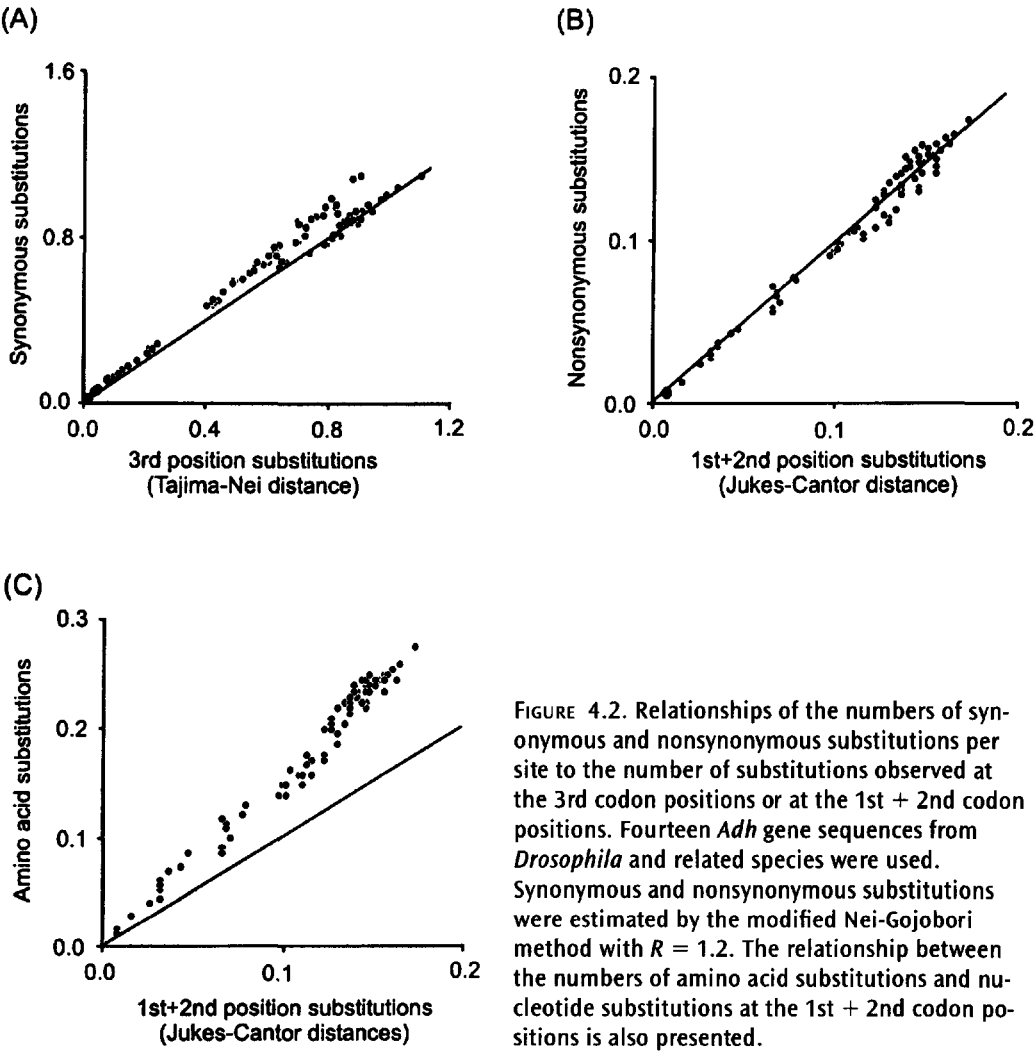

FIGURE 4.2. Relationships of the numbers of synonymous and nonsynonymous substitutions per site to the number of substitutions observed at the 3rd codon positions or at the 1st + 2nd codon positions. Fourteen *Adh* gene sequences from *Drosophila* and related species were used. Synonymous and nonsynonymous substitutions were estimated by the modified Nei-Gojobori method with $R = 1.2$. The relationship between the numbers of amino acid substitutions and nucleotide substitutions at the 1st + 2nd codon positions is also presented.

because they are generally free from selection. However, as the number of substitutions increases, the accuracy of the estimates is expected to decline, because the assumptions used to estimate the number of synonymous substitutions are unlikely to hold for a long time. As mentioned earlier, synonymous and nonsynonymous sites are not fixed but vary with time. For this reason, some authors prefer to use the number of nucleotide substitutions at third codon positions for estimating evolutionary times. At these sites, a certain proportion of nucleotide substitutions are nonsynonymous, but the nucleotide sites are clearly defined and do not change with time. Therefore, the number of substitutions at third positions may be linearly related with evolutionary time.

In practice, the number of synonymous substitutions for a gene is generally greater than the number of third-position substitutions. Figure 4.2A shows the relationship between the number of synonymous substitutions ($\hat{d}_S$) obtained by the modified Nei-Gojobori method and the number of third-position substitutions ($\hat{d}_3$) for the alcohol dehydrogenase

(*Adh*) gene sequences from 14 different *Drosophila* species (see the book's website: *http://www.oup-usa.org/sc/*0195135857). The $\hat{d}_3$ values were obtained by Tajima and Nei's method, because the nucleotide frequencies at third codon positions are substantially different from 0.25. The results show that $\hat{d}_S$ is generally slightly higher than $\hat{d}_3$ as expected, but there is an approximately linear relationship between $\hat{d}_S$ and $\hat{d}_3$ for $\hat{d}_S < 0.8$. This result suggests that either $\hat{d}_S$ or $\hat{d}_3$ can be used for estimating divergence times as long as $d_S < 0.8$ in the present case. In fact, Thomas and Hunt (1993) used $\hat{d}_S$ for estimating the times of divergences of various *Drosophila* species whereas Russo et al. (1995) used $\hat{d}_3$, but their results were virtually the same.

Figure 4.2B shows the relationship between the number of nonsynonymous substitutions ($\hat{d}_N$) and the Jukes-Cantor distance for first and second codon positions ($\hat{d}_{12}$) for the same data set of *Adh* gene sequences. Here the $\hat{d}_N$ and $\hat{d}_{12}$ values are much smaller than the $\hat{d}_S$ and $\hat{d}_3$ values, but $\hat{d}_N$ and $\hat{d}_{12}$ are nearly equal to each other for all sequence comparisons. This indicates that for estimating divergence times either $\hat{d}_N$ or $\hat{d}_{12}$ can be used. Previously, we mentioned that the number of amino acid substitutions often gives a good estimate of divergence time. Figure 4.2C shows the relationship between the Poisson correction distance ($\hat{d}$) for amino acid sequence data and $\hat{d}_{12}$. Here again we can see a good linear relationship, although $\hat{d}$ is greater than $\hat{d}_{12}$ as expected.

4.4. Likelihood Methods with Codon Substitution Models

Goldman and Yang (1994) developed a likelihood method for estimating the rates of synonymous and nonsynonymous nucleotide substitution considering a nucleotide substitution model for 61 sense codons. (Three nonsense codons were eliminated.) Their model is somewhat similar to the Hasegawa-Kishino-Yano model (Table 3.2E) for nucleotide substitution. Let us consider a pair of sequences of C homologous codons and let π_j be the relative frequency of the j-th codon. They assumed that the instantaneous substitution rate (q_{ij}) from codon i to codon j ($i \neq j$) is given by the following equations.

$$q_{ij} = \begin{cases} 0, & \text{if nucleotide change occurs at two or more positions} \\ \pi_j, & \text{for synonymous transversion} \\ k\pi_j, & \text{for synonymous transition} \\ \omega\pi_j, & \text{for nonsynonymous transversion} \\ \omega k\pi_j, & \text{for nonsynonymous transition} \end{cases} \quad (4.12)$$

where k is the transition/transversion rate ratio and ω is the nonsynonymous/synonymous rate ratio. Here k may be written as α/β if the rates of transitional and transversional changes are α and β, respectively. Similarly, ω may be written as r_N/r_S if the rates of synonymous and

nonsynonymous changes are r_S and r_N, respectively. Therefore, if ω is the same for all codon pairs as assumed, it is possible to relate r_N/r_S to d_N/d_S.

There are 61 parameters for π_j, but if we assume that the codon frequencies are in equilibrium, they can be estimated by the observed codon frequencies when the number of codons used (C) is large. Therefore, the only parameters to be estimated are k and ω, and these parameters can be estimated by using the maximum likelihood method (Goldman and Yang 1994). When C is relatively small, however, this approach does not give a reliable estimate of π_j, because π_j is generally very small and thus the sampling error of the estimate of π_j is large. In this case, one may estimate π_j by a product of the observed nucleotide frequencies. In the present approach $\omega < 1$, $\omega = 1$, and $\omega > 1$ represent purifying selection, neutral evolution, and positive selection, respectively. Therefore, if the estimate ($\hat{\omega}$) of ω obtained from the data is significantly greater than 1, positive selection is suggested. Theoretically, this test can be done by using the likelihood ratio test.

Let $\ln L_2$ be the log *maximum likelihood* (ML) value when ω is estimated from the data and $\ln L_1$ be the ML value when $\omega = 1$ (null hypothesis) is assumed. The log likelihood ratio is then given by

$$LR = 2(\ln L_2 - \ln L_1) \tag{4.13}$$

When the numbers of synonymous and nonsynonymous substitutions are sufficiently large and the model used is appropriate, LR is approximately χ^2 distributed with one degree of freedom. Therefore, if $\hat{\omega} > 1$ and $LR \geq 3.84$, one may conclude that the rate of nonsynonymous substitution is significantly higher than that of synonymous substitution at the 5% level and that this is due to positive selection.

One advantage of this approach is that both the transition/transversion rate ratio (k) and the nonsynonymous/synonymous rate ratio (ω) can be estimated simultaneously if the model given in Equation (4.12) is satisfied. Therefore, there is no need to know R ($= 2k$) to estimate d_S and d_N as in the case of the modified Nei-Gojobori method.

However, there seem to be several problems with this approach. First, estimates of π_j's based on the observed frequencies would not be reliable when C is small as mentioned above. Estimation of π_j's by products of nucleotide frequencies would also be unreliable when the codon usage bias exists. Second, the assumption that ω is the same for all codon positions is quite unrealistic as is clear from the pattern of amino acid substitution discussed in chapter 1. This would make $\hat{\omega}$ substantially different from $\hat{d}_N/\hat{d}_S$, because the average of the ratio r_N/r_S is not the same as the ratio of the averages of r_N and r_S. Third, the assumption of independence of k and ω for every codon pair also would not be satisfied in actual data. Therefore, a more careful study is necessary about the effect of violation of the assumption on $\hat{\omega}$.

We have used this method (the default option of the computer program PAML by Yang [1999]) to compute the $\hat{d}_S$ and $\hat{d}_N$ values for the MHC and mitochondrial *Nd5* genes discussed earlier. The results are presented in Tables 4.3 and 4.4. In the extracellular region of the MHC gene, $\hat{d}_N$ is sim-

ilar to the $\hat{d}_N$ values obtained by the other methods, but $\hat{d}_S$ is more than two times higher than the values obtained by the other methods. In the case of the ARS of the human MHC *A* locus, $\hat{d}_N$ was about 1000 times higher than $\hat{d}_S$. In mitochondrial gene *Nd5* the $\hat{d}_N$ and $\hat{d}_S$ are similar to those of the modified Nei-Gojobori method.

Muse (1996) developed a similar likelihood method based on a different codon substitution model. In this method, codon frequencies are estimated by products of nucleotide frequencies, and no transition/transversion bias is assumed. Therefore, the number of parameters to be estimated is less than in the Goldman-Yang model. This method seems to give $\hat{d}_S$'s and $\hat{d}_N$'s similar to those of the Nei-Gojobori method when codon usage bias is small. When this bias and the transition/transversion bias are high, however, Muse's method is expected to give biased estimates.

As the computer technology develops, it is possible to use increasingly complicated mathematical models and conduct statistical analyses based on these models (e.g., Nielsen and Yang 1998). However, as the mathematical model becomes sophisticated, more parameters are required, and the underlying assumptions are likely to be violated quite often. A sophisticated model therefore may give biased estimates of the parameters. In contrast, the evolutionary pathway methods discussed earlier are based on the concept of parsimony analysis and are largely model free. Adaptive amino acid substitutions usually occur at some specific sites for functional reasons, and the pattern of the substitutions are likely to be different from the general pattern of amino acid substitution. Particularly, when d_S and d_N are large (say d_S, $d_N > 0.4$), these methods appear to give less reliable estimates than the simple evolutionary pathway methods, because there are many disturbing factors that affect the estimates of d_S and d_N (Tanaka and Nei 1989; Nei and Hughes 1992).

Another problem with the likelihood approach is the reliability of the likelihood ratio test. This test requires that the assumptions of the mathematical model used are satisfied with real data (Foutz and Srivastava 1977). Zhang (1999) has shown that in the test of evolutionary hypotheses this requirement is often violated and that in this case the test can be either too liberal or too conservative depending on the situation. Note also that the likelihood ratio test is a large-sample test, so that it may give erroneous conclusions when the numbers of synonymous and nonsynonymous substitutions are small. Therefore, caution is necessary in the application of this test.

This page intentionally left blank