J Mol Evol (1995) 41:345–352

JOURNAL OF MOLECULAR EVOLUTION



# Kinetics of Synonymous Codon Change for an Amino Acid of Arbitrary Degeneracy

**Otto G. Berg**

Department of Molecular Biology, BMC, Box 590, Uppsala University Biomedical Center, S-75124 Uppsala, Sweden

Received: 20 October 1994 / Accepted: 5 January 1993

**Abstract.** The kinetics of synonymous codon change and species divergence is described in a matrix formalism that is equally applicable to all levels of codon degeneracy and all levels of codon or nucleotide bias. Based on the formalism it is possible to calculate the sum of all the synonymous substitution rate constants from the observed sequence differences between two species. This sum, the relaxation rate, is equivalent to the LogDet transformation that has recently been proposed as a new measure of evolutionary distance (Lockhardt et al. *Mol. Biol. Evol. 11*(4): 605–612, 1994). The relationship between this measure and the average number of base changes per site ($K_s$) is discussed. The formalism is tested on some sets of simulated sequence divergence data.

**Key words:** Synonymous substitution rates — Genome evolution — LogDet transformation

## Introduction

The nucleotide sequence dissimilarity in the genomes of related species can be used to infer the phylogenetic distance between two species. The dissimilarity can also be used to estimate the various substitution rates of one base pair for another. Such an analysis requires a large set of sites where the kinetics of base change is very similar. The simplest situation to describe is that of neutral substitutions which do not influence the phenotype. In this case, the substitution rates will be equal to the mutation rate constants, which may vary relatively little between sites. Also synonymous substitutions in coding sequences are often treated as neutral. However, many organisms, particularly unicellular ones, frequently display a distinct codon preference that indicates that synonymous substitutions cannot be selectively neutral (Ikemura 1981). Nevertheless, synonymous differences still have properties that make them useful for a statistical study of sequence dissimilarity. This is because it can be argued that a particular synonymous base change for a particular amino acid has approximately the same kinetics regardless of where in the genome it occurs. As a consequence, by grouping together all synonymous differences for each amino acid separately, it is possible to get a large statistical data set even from the study of the difference between only two species. When the codon preference is different in different parts of the genome, one must consider the statistics of the synonymous substitutions in different subgroups of the genome. For example, in many unicellular organisms, codon bias varies with the expression level of the gene considered (Ikemura 1981). Similarly, in genomes where the GC content varies strongly across the chromosomes, it would be important to group together genes that are in the same background of GC content since this is likely to influence the effective substitution rates.

In contrast, the rate constants for nonsynonymous substitutions are expected to vary considerably from site to site depending on the function of the corresponding amino acids at each particular site. Thus, nonsynonymous divergence cannot easily be described by kinetic models that assume the same rate constant for all substitutions where a certain nucleotide is replaced by a certain

other. Nevertheless, the evolutionary distance between two lineages is frequently estimated based on such kinetic models applied to all nucleotide substitutions independently of what kind of sites they appear at. This corresponds to treating all sites as fourfold degenerate and most of the models described below have been applied in this way. Methods have been developed to account for a statistical distribution of rate constants from site to site (Nei and Gojobori 1986). To ensure the validity of using a kinetic scheme that can provide information about the individual rate constants, this study will focus on the synonymous substitutions.

There are a large number of models for the calculation of evolutionary distance based on the kinetics of base-pair change. (For a recent review, see Rodriguez et al. 1990.) Most often the distance is defined as the average number of changes per site. As it turns out, this distance measure is quite insensitive to the assumed underlying substitution kinetics (Rodriguez et al. 1990). For instance, the two-parameter model of Kimura (1980) assumes that all nucleotides are equiprobable; nevertheless the model can be used to estimate the average number of changes even for sites where the nucleotide bias is large. Similarly, the model of Tajima and Nei (1984), which accounts for nucleotide bias but is valid only for a very restricted set of substitution schemes, also gives good estimates for the distance. However, these models cannot be used to estimate the underlying substitution rate constants except in the restricted cases for which they are valid. These and other similar models are frequently used to estimate the number of nonsynonymous changes; while this may provide useful estimates for evolutionary distance, it is doubtful that the assumed kinetics has any meaning when rate constants differ drastically from site to site.

It is the substitution rate constant of one nucleotide for another, rather than the average rate of change, that is most directly related to the mutation rates and relative selective advantage between the two. In a separate communication (Berg and Martelius 1995) we considered the kinetic equations of synonymous base change for a two-fold degenerate amino acid, taking proper account of mutational or selectional nucleotide bias, and used the model to study the rates of base change in *Escherichia coli* and *Salmonella typhimurium.* This made it possible to infer not only the various substitution rate constants but also to separate out the selection pressure and the mutation rate constants. In the present communication, the kinetic model is extended to describe the kinetics of synonymous change for an amino acid of arbitrary degeneracy. Because of the multitude of possible synonymous substitution paths for higher degeneracy than two, it is not possible to estimate the substitution rate constants except in special cases. The model presented below assumes a general set of rate constants for the possible synonymous substitutions. It is therefore applicable to any set of genes that are under the same evolutionary pressure for codon or nucleotide bias. The model is similar to that presented by Rodriguez et al. (1990), but with some of their assumptions removed. One quantity that is directly accessible from the divergence data is the sum of all the rate constants in the substitution scheme. This turns out to be the same as the LogDet transformation that has recently been proposed as the best measure for evolutionary distance (Steel 1994; Lockhardt et al. 1994). The properties of this measure under various kinetic schemes will also be discussed below.

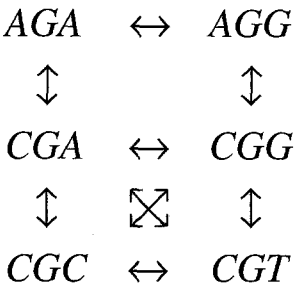


**Fig. 1.** The substitution scheme for the six codons of arginine. The codons can be numbered and each possible nucleotide substitution (*arrows* in the scheme) can be assigned a rate constant as described in the text.

## Kinetics for Synonymous Codon Change

Let us consider the synonymous codon choices for an $n$-fold degenerate amino acid. The codons can be numbered $i = 1,2, \ldots, n$. Assume that the substitution rate constant from codon $i$ to codon $j$ is $k_{ij}$. These are called rate constants in analogy with chemical kinetic equations in order to distinguish them from the average rate of change that will also be considered below. For a sixfold degenerate amino acid (see Fig. 1) not all codon choices can be connected via a single nucleotide substitution; these probably have very small and possibly zero values for the corresponding value of $k_{ij}$. Thus the rate constants that enter the scheme are based on single nucleotide replacements. The rate constants define a matrix, **k,** with the elements $k_{ij}$ for $i \neq j$ and the diagonal elements equal to $k_{ii} = -\sum_{j \neq i} k_{ij}$. Thus, the elements of each row in the matrix will sum to zero. If $P_i(t)$ denotes the probability that a certain site uses $i$ at time $t$, the vector of probabilities $\mathbf{P}(t) = \{P_i(t)\}$ will satisfy the rate equation

$$\frac{d\mathbf{P}}{dt} = \mathbf{P} \cdot \mathbf{k} \tag{1}$$

The formal solution to the rate equation is

$$\mathbf{P}(t) = \mathbf{P}(0) \cdot \exp(\mathbf{k}t) \tag{2}$$

Since the sums and products of matrices are well defined,

the time evolution matrix can be defined formally by the series expansion

$$\mathbf{T}(t) = \exp(\mathbf{k}t) = \sum_{n=0}^{\infty} \frac{1}{n!} \mathbf{k}^n t^n \qquad (3)$$

Thus, if the substitution rate constants and the initial conditions are known, the time evolution of the probabilities for each synonymous codon choice is determined. The elements of the time evolution matrix are

$$T_{ij} = p_j(t|i) \qquad (4)$$

which is the probability for codon $j$ at time $t$ conditional on there being a codon $i$ at time $t = 0$.

## Accumulation of Differences Between Two Species

Consider two species that separated from a common ancestor at time $t = 0$. The divergence between the two species at some time $t$ later can be described by a dissimilarity matrix, **D.** This matrix has the elements $D_{ij}$ given as the fraction of sites for a particular amino acid where species 1 has codon $i$ *and* species 2 has codon $j$. At time $t = 0$, when the two species are equal to their common ancestor, the dissimilarity matrix is diagonal and determined from the initial condition: $D_{ij}(0) = 0$ for $i \neq j$ and $D_{ii}(0) = P_i(0)$. Since changes in the two species are independent, we can apply the kinetic relation, eq. (2), for each and find

$$\mathbf{D}(t) = \mathbf{T}_1^T \cdot \mathbf{D}(0) \cdot \mathbf{T}_2 = \exp(\boldsymbol{k}_1^T) \cdot \mathbf{D}(0) \cdot \exp(\mathbf{k}_2 t) \quad (5)$$

$\mathbf{k}_1$ is the substitution rate matrix in species 1 and $\mathbf{k}_2$ the corresponding expression for species 2. $\mathbf{k}^T$ is the transposed matrix where the elements $k_{ij}$ have been replaced by $k_{ji}$. Similarly, the corresponding time evolution matrices $\mathbf{T}_1$ and $\mathbf{T}_2$ are given by eq. (3).

For example, from eq. (5) the element $D_{ij}$ of the dissimilarity matrix is determined by

$$D_{ij}(t) = \sum_m P_m(0) P_i^{(1)}(t|m) P_j^{(2)}(t|m) \qquad (6)$$

The probabilistic structure of this expression is obvious: The first term is the product of the probability $P_m(0)$ that there was codon $m$ at time 0, the probability $P_i^{(1)}(t|m)$ that species 1 has codon $i$ at time $t$ conditional on there being codon $m$ at time 0, and the probability $P_j^{(2)}(t|m)$ that species 2 has codon $j$ at time $t$ conditional on there being codon $m$ at time 0. Thus as it should, the sum of all terms in eq. (6) expresses the total probability that there is codon $i$ in species 1 simultaneously as there is codon $j$ in species 2 at time $t$.

Eq. (5) is the general relation expressing the divergence between the two species at time $t$. In a sequence comparison, all the elements of **D** can be calculated from the observed differences and similarities. However, it is not possible in general to calculate the substitution rate matrices $\mathbf{k}_1$ and $\mathbf{k}_2$, even if the original base distribution at the time of separation, **D**(0), were also known; there is not enough information in the result from two time points, the initial and the present, to infer all rate constants. However, by taking the determinant of both sides of eq. (5), the sum of all the substitution rate constants can be calculated directly from the observed dissimilarity matrix:

$$R_S = \sum_{\substack{i,j \\ i \neq j}} (k_{ij}^{(1)} + \mathrm{k}_{ij}^{(2)} t) = -\ln\left(\frac{\det(\mathbf{D}(t))}{\det(\mathbf{D}(0))}\right) \qquad (7)$$

where $k_{ij}^{(1)}$ is the rate constant of replacement of codon $i$ to $j$ in species 1 and correspondingly for species 2. This result follows from a number of properties of the matrices involved: (1) The determinant of a product equals the product of the determinants; (2) the determinant of the time evolution matrix, $\mathbf{T} = \exp(\mathbf{k}t)$, is the product of its eigenvalues; (3) this product equals the exponential of the sum of eigenvalues of $\mathbf{k}t$; (4) the sum of the eigenvalues for **k** equals the sum of all its rate constants. When, after a long separation time, the synonymous differences become saturated, det(**D**) approaches zero and eq. (7) is useless; in this limit the synonymous substitutions carry little information.

The proper calculation of $R_s$ requires a knowledge of the initial condition, **D**(0). The determinant of the initial condition is just the product of the initial four base probabilities. If both lineages develop with the same codon distribution, the most reasonable assumption is that their common ancestor also had this distribution. Even if the two lineages develop towards different codon distributions, the choice of **D**(0) is not crucial as long as the difference in codon bias is not dramatic; it can with relatively small errors (see below) be taken as the average of the distributions in the two.

Since this expression is proportional to the separation time, $t$, eq. (7) may provide the most useful measure for phylogenetic distance. This LogDet expression has also been derived without recourse to a kinetic scheme and suggested to provide the correct evolutionary distance (Steel 1994; Lockhardt et al. 1994). However, most often one uses the average number of base substitutions per site as a measure of the evolutionary distance. This cannot be calculated directly from the data without introducing additional assumptions about the underlying kinetics, as will be evident below.

Based on eq. (5), an equivalent way of calculating the sum of all rate constants is:

$$R_s = -\mathrm{trace}[\ln(\mathbf{D}^{-1}(0) \cdot \mathbf{D}(t))] \qquad (8)$$

where $\mathbf{D}^{-1}(0)$ is the inverse of the initial condition ma-

trix. This expression is much less useful than eq. (7) but is given here for comparison with eq. (14) below. The logarithm of the matrix can be calculated either as described, e.g., by Rodriguez et al. (1990), or from the series expansion

$$\ln(\mathbf{D}^{-1}(0)\cdot\mathbf{D}(t)) = -\sum_{n=1}^{\infty}\frac{1}{n}(\mathbf{I}-\mathbf{D}^{-1}(0)\cdot\mathbf{D}(t))^n \quad (9)$$

where $\mathbf{I}$ is the identity matrix. The sum can easily be calculated on a computer to any required degree of accuracy through a finite number of terms.

## Equilibrium Relaxation

In one important special case, the relation (5) can be considerably simplified so that all the substitution rate constants can be calculated directly from the dissimilarity matrix. Assume that the two species have evolved since separation with the same codon bias $P_i^o$ as their common ancestor had. Then the observed codon frequencies in both species would be the same and the same as the initial probabilities $P_i(0)$. The codon distribution in both organisms remains the same and the divergence is due only to random switches within this distribution.

One more assumption is required before a solution for the rate constants can be achieved. This is the reversibility condition (cf. Rodriguez et al. 1990):

$$k_{ij}P_i^o = k_{ji}P_j^o \qquad for\ i \neq j \quad (10)$$

which corresponds to the principle of detailed balance for a thermodynamic equilibrium: For every possible pair of substitutions $(i,j)$, the number of changes of $i$ for $j$ equals the number of changes of $j$ for $i$. If this relation holds, the matrix $\mathbf{k}^T\mathbf{D}_0$ will be symmetric:

$$\mathbf{k}^T\cdot\mathbf{D}_0 = (\mathbf{k}^T\cdot\mathbf{D}_0)^T = \mathbf{D}_0\cdot\mathbf{k} \quad (11)$$

The first equality is the symmetry relation and the second follows because $\mathbf{D}_0$ is diagonal. If both $\mathbf{k}_1$ and $\mathbf{k}_2$ satisfy eqs. (10) and (11), they will commute and eq. (5) can be written as

$$(\mathbf{k}_1+\mathbf{k}_2)t = \ln(\mathbf{D}_0^{-1}\cdot\mathbf{D}(t)) \quad (12)$$

This expression is valid only for kinetic models that produce a divergence matrix $\mathbf{D}(t)$ which is symmetric. Conversely, only a symmetric $\mathbf{D}$ in eq. (12) will produce a rate constant matrix where the rows sum to zero as required. Small-sample variations may produce a nonsymmetric divergence matrix even if the underlying kinetics satisfy the conditions for eq. (12). In such cases it is best to symmetrize $\mathbf{D}$ by replacing it with $0.5(\mathbf{D}+\mathbf{D}^T)$ before use in eq. (12).

The average number of synonymous changes per site in the two species during the separation time $t$ is the weighted average

$$K_s = \sum_{\substack{i,j\\ i\neq j}}(k_{ij}^{(1)}+k_{ij}^{(2)})P_i^0 t \quad (13)$$

This can be calculated from the solution, eq. (12), as

$$\begin{aligned} K_s &= -\text{trace}(\mathbf{D}_0\cdot(\mathbf{k}_1+\mathbf{k}_2)t) \\ &= -\text{trace}[\mathbf{D}_0\cdot\ln(\mathbf{D}_0^{-1}\cdot\mathbf{D}(t))] \end{aligned} \quad (14)$$

This result has been derived previously by Rodriguez et al. (1990). Since the logarithm cannot be calculated when $\det(\mathbf{D}) \leq 0$, this relation for $K_s$ becomes inapplicable in the same cases as eqs. (7) and (8) for $R_s$ do.

The mutational appearance and fixation of base substitutions is a driven process far from thermodynamic equilibrium. Thus there is no reason to expect detailed balance to hold for this system, even if it is "equilibrated" in the sense that codon distributions remain constant. Nevertheless, eq. (12) may provide the best estimates for the substitution rate constants from the limited amount of data available. The rate constants calculated from eq. (12) also satisfy the exact relation, eq. (7), for which no extra assumptions were needed.

The relations derived above are valid for all synonymous substitutions, be they 2-, 3-, 4-, or 6-fold degenerate, with an obvious choice of dimensions for the matrices involved. In the twofold degenerate case, where detailed balance is automatically satisfied, the result agrees with our previous one (Berg and Martelius 1995). Eq. (14) is valid under equilibrium relaxation for arbitrary substitution rate matrices that satisfy the detailed balance condition; it may be the most general expression for $K_s$ that can be found (Rodriguez et al. 1990). Below, we shall compare this with the results from some more commonly used models which are valid under much more restricted substitution schemes. The two-parameter model of Kimura (1980) assumes that all transitions take place with one rate constant and all transversions with another. As a consequence, it leads to equal frequencies for all nucleotides and it may therefore be expected to be less useful for situations with strong nucleotide preferences or codon bias. The formulation of Tajima and Nei (1984) is strictly valid only for the equal-input model where the rate constant to a particular nucleotide is independent of which nucleotide is being replaced. This model gives a very simple expression for the average number of changes per site:

$$K_s = -b\ln(1-p/b) \quad (15)$$

where

$$p = 1 - \text{trace}\,(\mathbf{D}) \quad (16)$$

is the fraction of sites with differences in the two lineages, and

$$b = 1 - \sum_i (P_i^0)^2 \qquad (17a)$$

is the fraction of differences expected after infinite time. If divergence takes place under nonequilibrium conditions such that there are different codon frequencies, $P_i^{(1)}$ and $P_i^{(2)}$, in the two lineages, this relation can be changed to (Bulmer 1991)

$$b = 1 - \sum_i P_i^{(1)} P_i^{(2)} \qquad (17b)$$

As we shall see below, eq. (17b) provides a useful approximation for nonequilibrium relaxation. Tajima (1993) has extended both the Tajima and Nei (1984) model and the two-parameter model (Kimura 1980) to provide unbiased estimates also for a small sample of sites.

## Results

The equations allow us to calculate the expected synonymous sequence divergence for any sets of substitution rate constants. This can be done either from the deterministic relation, eq. (5), which would be valid for an infinite number of sites, or it can be done by simulation for a finite number of sites. From the dissimilarity matrices generated in this way, one can calculate the $K_s$ and $R_s$ values based on the different models and compare with the appropriate averages of the rates that went into the calculation. There are two aspects to consider: First, there are the systematic differences that appear because various assumptions of the kinetic models are not satisfied; these can be studied using the deterministic equations. Second, one must consider the applicability of the equations to divergence matrices than contain small-sample variations from the use of a finite number of sites; this can most easily be studied by simulations. The probability for a particular change from codon $i$ to codon $j$ in genome 1 and 2 at any time is $k_{ij}^{(1)} P_i^{(1)}$ and $k_{ij}^{(2)} P_i^{(2)}$, respectively; $P_i^{(1)}$ is the fraction of sites in genome 1 with codon $i$ at the time considered, and correspondingly for genome 2. In the simulations described below, each successive change was assigned with this probability and the codon frequencies were adjusted after each. An initial codon distribution was assumed for each simulation, but since this is not usually known in general, for the analysis of divergence using eqs. (7) or (14) an initial distribution, $\mathbf{D}(0)$, was estimated by taking the average of the observed frequencies of the different codons in the two species.

Rodriguez et al. (1990) simulated the divergence for fourfold degenerate sites using various different rate matrices under equilibrium relaxation. They found that the distances estimated from the $K_s$ values were quite close to their real value regardless of which underlying kinetic model was used and regardless of whether the conditions assumed for the different models were satisfied or not. I have tested both equilibrium and nonequilibrium divergence, but using a more severe bias for the nucleotide choices. Figure 2 shows the results for the equilibrium relaxation of fourfold degenerate sites when two nucleotides (e.g., the purines) are tenfold preferred over the

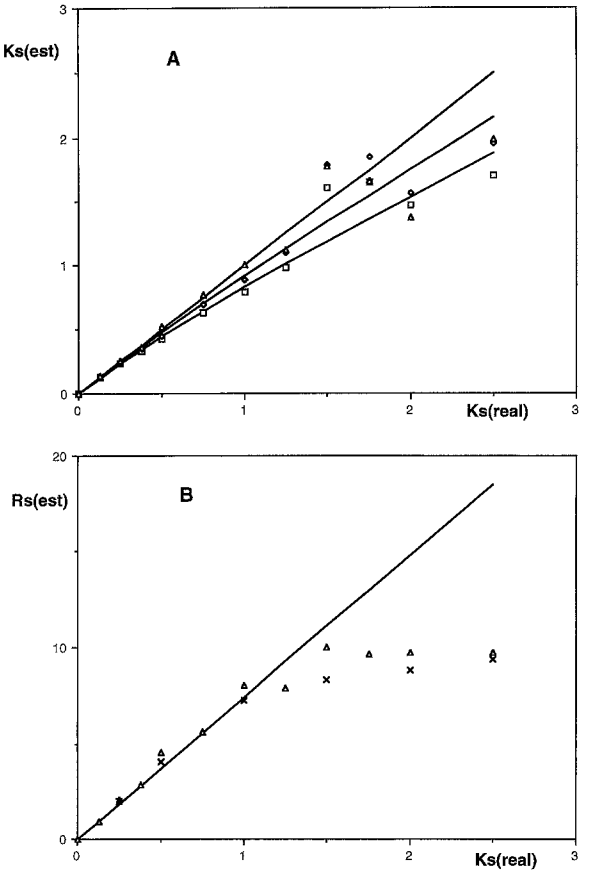


**Fig. 2.** Equilibrium relaxation for fourfold degenerate codons where two codons are ten times more common than the other two. The substitution rate constant matrices used in the calculations were

$$\mathbf{k}_1 = \mathbf{k}_2 = \begin{pmatrix} -1.4 & 1 & 0.2 & 0.2 \\ 1 & -1.4 & 0.2 & 0.2 \\ 2 & 2 & -5 & 1 \\ 2 & 2 & 1 & -5 \end{pmatrix}$$

**A** shows the estimated $K_s$ values after various times of separation using the two-parameter model (*lowest line and squares*), the Tajima and Nei model (*middle line and diamonds*), and eq. (14) (*upper line and triangles*). The *lines* are for infinite number of sites and the *data points* are from one simulation using 400 sites. The exact result would be according to the *upper straight line,* where the estimated number of changes equals the real number. **B** shows the estimated relaxation rate ($R_s$) from eq. (7). The *straight line* is the exact result for an infinite number of sites. The *diamonds* are from one simulation using 400 sites and the x's are the averages from 100 such simulations.

other two. Based on the deterministic equations (infinite number of sites), Kimura's two-parameter model systematically underestimates the true value for $K_S$, while the estimate based on the Tajima and Nei model is closer to the true one. The estimate based on eq. (14) is exact under these conditions, as is the estimate for $R_s$ from eq. (7). The differences between the models become somewhat blurred in the simulations of a finite number of sites. In the simulations, the estimated $K_s$ and $R_s$ values do not continue to increase linearly in time as expected from the deterministic equations, but level off the fluctuate around some finite value. At this point the divergence is saturated and distances cannot be estimated from any model. The smaller the number of sites studied, the earlier saturation occurs.

For an infinite number of sites under equilibrium relaxation, the relaxation rate $R_s$ from eq. (7) would provide the best distance estimate since the expression is linear in the divergence time and exact for all possible substitution schemes. The estimate of $K_s$ from eq. (14) is also exact under these circumstances, but only if the substitution scheme also satisfies the detailed balance condition. For a finite number of sites, however, both estimates are sensitive to small-sample variation and saturation effects. For some of the time points, the simulations using 400 sites have been repeated 100 times. After an average of 0.75 changes per site, 23% of the runs give a negative value for det(**D**), which therefore give inapplicable results for the use of eqs. (7) and (14). After an average of one and 1.5 changes per site there are 40% and 53% inapplicable cases, respectively. That is, above one change per site, the results for $R_s$ are virtually randomized with det(**D**) fluctuating around zero. The Tajima and Nei estimate, in contrast, is much less sensitive to the small-sample saturation effects and provides 100% applicable cases for these time points with a significant reduction only after an average of two changes per site when there are 89% applicable results. Below saturation, the relative scatter is around 25% in the $R_s$ estimates while it is less than 10% in the $K_s$ estimate from Tajima and Nei. Thus, for the particular substitution scheme studied in Fig. 2, the most useful distance estimate may be provided by the approximate Tajima and Nei estimate, and not by the exact relation for $R_s$ in eq. (7).

In Fig. 3 are displayed the results for a nonequilibrium relaxation where the two lineages start at the same strong codon bias, but after separation the codon bias in one of the lineages is assumed to relax toward equal frequencies of all codons. For this situation, none of the $K_s$ estimates discussed above is exact. In this case it is the estimate based on Kimura's two-parameter model that comes closest to the real number of changes per site. The Tajima and Nei estimate works fairly well up to about 1.5 changes per site; in a run of 100 different simulations there are a 100% applicable cases and 10% relative scatter at this point and below. The $R_s$ values based on the same data are more variable and become very unreliable

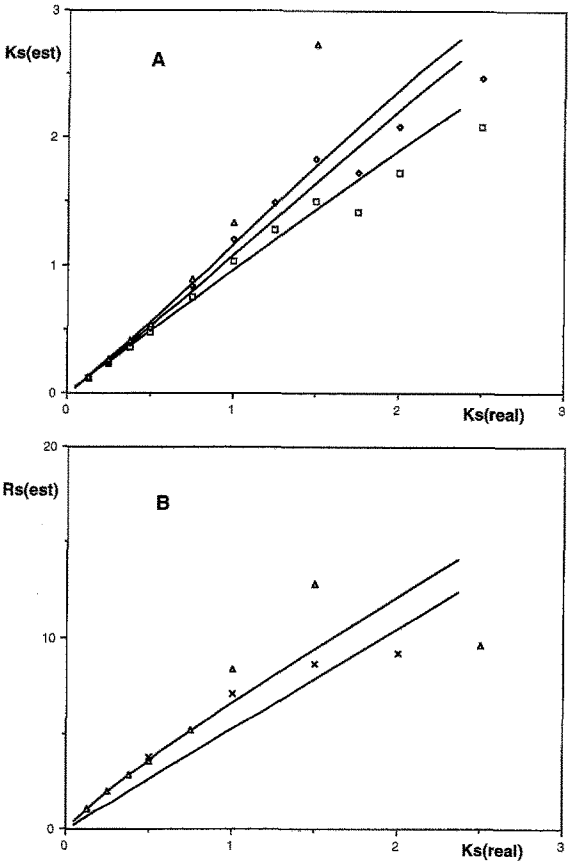


**Fig. 3.** Nonequilibrium relaxation for fourfold degenerate codons where two codons are ten times more common than the other two in one of the lineages and in the ancestor. The other lineage relaxes toward an equal codon distribution. The substitution rate constant matrices used in the calculations were

$$\mathbf{k}_1 = \begin{pmatrix} -1.4 & 1 & 0.2 & 0.2 \\ 1 & -1.4 & 0.2 & 0.2 \\ 2 & 2 & -5 & 1 \\ 2 & 2 & 1 & -5 \end{pmatrix} \text{ and } \mathbf{k}_2 = \begin{pmatrix} -3 & 1 & 1 & 1 \\ 1 & -3 & 1 & 1 \\ 1 & 1 & -3 & 1 \\ 1 & 1 & 1 & -3 \end{pmatrix}$$

**A** shows the estimated $K_s$ values after various times of separation using the two-parameter model (*lowest line and squares*), the Tajima and Nei model (*middle line and diamonds*), and eq. (14) (*upper line and triangles*). The *lines* are for infinite number of sites and the *data points* are from one simulation using 400 sites. **B** shows the estimated relaxation rate ($R_s$) from eq. (7). The *lower straight line* is the exact result for an infinite number of sites if the ancestral codon distribution is known. The *upper line* shows the same for the case where the ancestral codon distribution has been estimated from the average observed in the two lineages. The diamonds are from one simulation using 400 sites and the x's are the averages from 100 such simulations.

already around an average of one change per site. The estimate based on eq. (7) would be exact also under these conditions, if the codon distribution, **D**(0), in the ancestor were known; this is the straight line in the lower panel of Fig. 3. However, if the ancestral codon distribution is estimated as the average of the two lineages, one gets the upper line for an infinite number of sites. The data points are from a simulation using 400 sites. For a run of 100

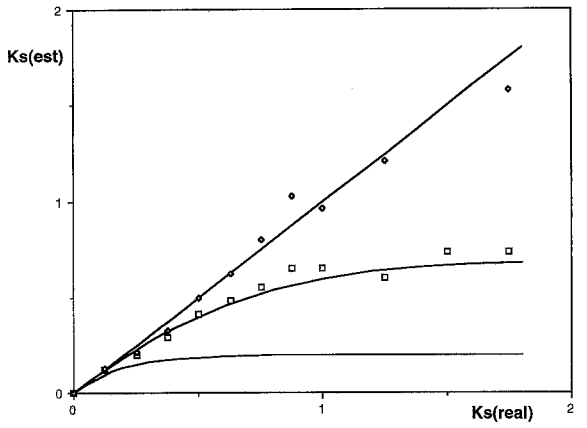


**Fig. 4.** Equilibrium relaxation of twofold degenerate sites. The *upper line* is the exact result (regardless of bias) which follows from use of eq. (14) or (15). The *middle line* is from an application of the two-parameter model for the case when one codon is preferred by a factor of three over the other. The *lower line* is the same if the codon preference is a factor of ten. The data points are from one simulation of 400 sites with bias three; *squares* are from an application of the two-state model to the data and *diamonds* are from use of eqs. (14) or (15).

simulations, one finds 21% inapplicable cases and a 20% relative scatter in the $R_s$ estimates after an average of one change per site.

Because of the statistical fluctuations, it does not appear useful to calculate the individual rate constants from eq. (12), except for samples including a very large number of sites (around 1,000 or more in the fourfold degenerate case). However, the different averages, $K_s$ and $R_s$, plotted in Figs. 2 and 3, are useful and readily available from the data. I have simulated (data not shown) also cases where the substitution-rate constants are very disparate and far from a detailed-balance condition. Such extra complications do not seem to change the general picture very much from that displayed in the figures above.

For twofold degenerate codons with or without bias and under equilibrium relaxation, both $K_s$ estimates based on eq. (14) and on the Tajima and Nei model, eq. (15), are exact, as is the $R_s$ estimate from eq. (7). The $K_s$ estimate based on Kimura's two-parameter model, on the other hand, departs more strongly than in the fourfold degenerate case. The lines in Fig. 4 show the results for the equilibrium relaxation of an infinite number of sites. The upper straight line is the exact result that can be calculated either from eq. (14) or eq. (15). This would be the result from the two-parameter model only if the two codons are equal in frequency (no bias). The two curved lines are from an application of the two-parameter model to two cases with strong codon bias which obviously would lead to serious underestimates of the average number of synonymous codon changes.

For sixfold degenerate codons (Fig. 1), neither the two-parameter model nor the Tajima and Nei model can be applied directly without corrections for the structure of the substitution scheme. Eq. (7) for $R_s$, however, gives directly the sum of all the rate constants in the scheme; for the scheme in Fig. 1 there are 18 rate constants involved. Similarly, under equilibrium and detailed balance, eq. (14) gives directly the average number of synonymous substitutions per amino acid involved.

If sites of very different kinetics, e.g., nonsynonymous and synonymous, are lumped together in the calculation of the dissimilarity matrix, none of the measures discussed above will give a reasonable estimate for the average rate of change, except possibly at early times after separation when the divergence is due only to the fast substitutions. At later times, when the fast substitutions are saturated, they will only contribute a constant term to the estimated rate of change ($R_s$ or $K_s$), although they will continue to change at a rapid rate. Consequently, in this case $R_s$ or $K_s$ used as distance measure will severely underestimate the longer separation times compared to the shorter ones. They may still provide useful measures for evolutionary distance, but would say very little about the actual rates of change or average number of changes.

## Discussion

To understand the kinetics of base-pair substitutions, it is important to consider only sites that can reasonably be assumed to have the same rate constants for the same kind of substitutions. By lumping together sites that have very different substitution kinetics one cannot hope to get very reliable estimates even for averages of the rate constants. Instead, it is suggested that one study the synonymous substitutions separately for each amino acid. Possibly one can lump together amino acids of the same degeneracy if they have similar codon bias. Then, using eq. (7), one can calculate the sum of all the synonymous rate constants for the amino acid considered. An average synonymous rate constant for that amino acid would then be given by $R_s$ divided by the total number of rate constants involved in the substitution scheme (e.g., 18 rate constants for the scheme in Fig. 1); this normalization puts the rates for all amino acids on the same base and an overall average can be calculated by taking the amino-acid weighted average across the gene(s) studied. Similarly, $K_s$ estimated from eq. (14) will give an average number of synonymous changes for each amino acid and the overall can be synonymous changes for each amino acid and the overall average can be calculated by taking the amino-acid weighted average. These results can be used also for estimating the evolutionary distance between two lineages and—for the reasons discussed above—they are expected to be more closely linear in the separation time than measures based on all sites lumped together.

The divergence at synonymous sites allows a straightforward statistical analysis based on kinetic schemes, but due account must be taken of the biases involved. The

synonymous substitutions are useful only for species that are sufficiently closely related so that the differences have not reached saturation. Also, after long separation times it is not reasonable to consider the synonymous changes independently from the nonsynonymous ones. In a separate communication (Berg and Martelius 1995) we derived the relations for the twofold degenerate amino acids and used them to study the synonymous substitutions for the twofold degenerate amino acids in *E. coli* and *S. typhimurium*. By focusing on the individual substitution rate constants rather than on their weighted average, it was possible to separate mutation and selection pressure and also study the relationship between selection pressure and substitution rates. We intend to use the present model to extend this study to amino acids of higher degeneracy.

The matrix formulation has the advantage that it automatically accounts for all amino-acid degeneracies, 2-, 3-, 4-, or 6-fold, and thereby puts all synonymous changes on the same basis. The data required are in the dissimilarity matrix, counting the fraction of sites for a certain amino acid that uses a certain codon in one species and another (or the same) in the other species. Furthermore, for use with eq. (7) or (14), one needs an estimate of the codon preferences in the ancestor of the two species. This could be chosen as the average of the codon choices in the two species studied, at least if the difference in bias is not large. If several lineages are studied, one might find other evolutionary arguments to infer a most likely codon distribution in the ancestor(s). If this can be reliably estimated, the relaxation rate $R_s$ from eq. (7) is the only measure that remains exact also under nonequilibrium relaxation when two species drift toward different codon or nucleotide distributions. With additional assumptions, the individual substitution rate constants can also be directly calculated—eq. (12).

The measure for evolutionary distance has traditionally been the average number of changes per site that has taken place during the separation time; this has an obvious intuitive appeal as a measure for evolutionary change. For synonymous differences that is given by the weighted average, $K_s$, in eq. (13) above. This is directly available from the dissimilarity data via eq. (14) only after an assumption about constant codon bias and detailed balance. It gives a weighted average of the rate constants. Both this average and the relaxation rate, $R_s$ from eq. (7), are proportional to separation time and could therefore be used as a distance measure. Since $R_s$ can be calculated directly from the dissimilarity matrix with no additional assumptions, it may provide a better estimate for the evolutionary distance. However, for the particular kinetic schemes studied in Figs. 2 and 3, it turns out that the various estimates for $K_s$—in spite of their limitations—can provide estimates that are less sensitive to small-sample effects and saturation. However, for the twofold degenerate sites with bias, Fig. 4, the two-parameter model (Kimura 1980) cannot be applied except at very short separation times.

*Acknowledgment.* This work was supported by the Swedish Natural Science Research Council.

## Note Added in Proof

After this paper had been submitted, there appeared an extensive study of the relationship between the estimated phylogenetic distance and assumed substitution model (Zharkikh 1994).

## References

Berg OG, Martelius M (1995) Synonymous substitution rate constants in *Escherichia coli* and *Salmonella typhimurium* and their relationship with gene expression and selection pressure. J Mol Evol (in press)

Bulmer M (1991) Use of the method of generalized least squares in reconstructing phylogenies from sequence data. Mol Biol Evol 8: 868–883

Ikemura T (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. J Mol Biol 146:1–21

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

Lockhardt PJ, Steel MA, Hendy MD, Penny D (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. Mol Biol Evol 11:605–612

Nei M, Gojobori T (1986) Simple method for estimating the number of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418–426

Rodríguez F, Oliver JL, Marín A, Medina JR (1990) The general stochastic model of nucleotide substitution. J Theor Biol 142:485–501

Steel M (1994) Recovering a tree from the leaf colourations it generates under a Markov model. Appl Math Lett 7;19–23

Tajima F (1993) Unbiased estimation of evolutionary distances between nucleotide sequences. Mol Biol Evol 10:677–688

Tajima F, Nei M (1984) Estimation of evolutionary distances between nucleotide sequences. Mol Biol Evol 1:269–285

Zharkikh A (1994) Estimation of evolutionary distances between nucleotide sequences. J Mol Evol 39:315–329