Inferring the Number of Evolutionary Events from DNA Coding Sequence Differences

R. C. Lewontin
Museum of Comparative Zoology, Harvard University

The estimation of the amount of evolutionary divergence that has taken place between two DNA coding sequences depends strongly on the degree of constraint on amino acid replacements. If amino acid replacements are relatively unconstrained, the individual nucleotide is the appropriate unit of analysis and the method of Tajima and Nei can be used. If amino acid replacements are constrained, however, this method is shown to be inapplicable. For sequences with strong amino acid constraints, a method is outlined analogous to the Tajima and Nei method using codons as the unit of analysis. Only synonymous substitutions are used. Codon usage data can be employed to estimate the necessary parameters of the calculation, or a priori models of substitution may be employed. Sequences with significant but intermediate constraints on amino acid replacements are, in principle, unanalyzable.

Introduction

It is obvious that DNA sequence differences between organisms contain a great deal of information about how much evolutionary divergence has occurred since these organisms had a common ancestor. As the mass of sequence information in different species has accumulated, a good deal of attention has been devoted to methods of inferring evolutionary events from observed sequence comparisons (see Nei 1987, pp. 64–79, for a thorough review of the present knowledge).

DNA sequence data have two special properties that make the problem of inference different from those in the usual phylogenetic methods. First, any nucleotide position has only four possible states, so that, although two sequences may be observed to be identical at some position, they may be separated by numerous evolutionary changes but have converged to the same state. Second, each state is, at least mutationally, accessible from each other state, so sequences of evolutionary states are not well ordered. There is no way of knowing which states were intermediate between an initial and final one in an evolutionary pathway.

These properties, then, give rise to a problem of inference, the multiplicity problem. Two compared nucleotide states may be identical yet separated by 0, 2, 3, \ldots n substitution events, or, if they are different, they may be separated by 1, 2, 3 \ldots n events. Thus, the observed proportion of difference between two sequences is an underestimate of the number of evolutionary substitutions, and, even if an infinite number of substitutional events have occurred, the segments will be identical at some proportion of their positions. All methods of analysis have had, as their primary purpose, the production of an estimate of evolutionary divergence that takes into account the non-
linear relation between the observed proportion of sequence differences and the number of evolutionary substitutions.

The solutions to this inference problem have taken as their basic unit of analysis the individual nucleotide. A nucleotide sequence is regarded as a collection of sites each of which can exist in four states and each of which, if it is in state $i$, has some probability per unit time, $\pi_{ij}$, of an evolutionary substitution to state $j$. The fundamental assumption is that all nucleotide positions have the same set of $\pi_{ij}$'s irrespective of their actual location in the sequence. That is, there is some probability per unit time of passing from A to T, from C to G, from C to A, etc., which is the same at all positions. If we knew these probabilities and the paired nucleotide comparisons between two sequences, we could estimate how many actual substitution events had occurred between the sequences.

An Overview of the Theory

The solution to the problem of inference in the nucleotide model is completely known in principle and, to an extremely good order of approximation, in practice (see Nei 1987, pp. 64-79, for a discussion of many aspects of the problem). If all four nucleotides are equally likely to be substituted by any of the other three, we shall call this the isotropic model. Let $d =$ average number of evolutionary substitutions per nucleotide pair, between two sequences, and let $P =$ observed proportion of sites at which the two sequences differ. Then, by a differential equation approach, Jukes and Cantor (1969) showed that

$$d = -\frac{3}{4} \ln (1 - \frac{4}{3}P).$$

Using the same approach, Farris (1973) and Rice (1987) generalized (1) by showing that if there are $n$ isotropically related alternatives, then

$$d = \frac{n - 1}{n} \ln \left(1 - \frac{n}{n-1}P\right).$$

We note that $\frac{n - 1}{n}$ is the saturation differentiation between two sequences that will be approached as more and more evolutionary events occur between the sequences.

If, in contrast, the various substitutions of nucleotides have generally unequal probabilities, the Jukes-Cantor-Rice approach must be replaced by a more general substitution matrix method. Let $\pi_{ij}$ be the probability that after one unit of time (some small number of generations) a nucleotide in state $i$ has been substituted by state $j$. So $\pi_{ii}$ is the probability of no substitutions in this time interval. Let

$$\Pi_i = 1 - \pi_{ii} = \sum_j \pi_{ij} (i \neq j)$$

and

$$g_i = \text{the frequency of the } t^{th} \text{ base in the original sequence.}$$

Then, after $t$ units of time

$$g^{(t)} = M^t g,$$
where $g$ and $g^{(t)}$ are column vectors of the $g_i$ initially and after $t$ time units, respectively, and where $M^t$ is the $t$th power of the column-wise stochastic matrix

$$M = \begin{bmatrix} 1 - \Pi_1 & \pi_{21} & \pi_{31} & \pi_{41} \\ \pi_{12} & 1 - \Pi_2 & \pi_{32} & \pi_{42} \\ \pi_{13} & \pi_{23} & 1 - \Pi_3 & \pi_{43} \\ \pi_{14} & \pi_{24} & \pi_{34} & 1 - \Pi_4 \end{bmatrix}. \quad (4)$$

It then follows that after $t$ units of time, the proportion $P$ of nucleotide positions that have an observed difference is given by

$$P = \sum_{i=1}^{4} g_i \Pi_i^{(t)}, \quad (5)$$

where $1 - \Pi_i^{(t)}$ is the $i$th element in the principal diagonal of $M^t$.

The solution to the inference problem is then a two-step process:

1. Raise the matrix to successively higher powers and try relation (5) until it is satisfied. This gives an estimate of the time, $t$.
2. The estimated total number of evolutionary events $d$ is related to the time by the simple linear relationship

$$d = \sum_{i=1}^{4} g_i \Pi_i^{t}. \quad (6)$$

At first sight this solution seems impractical. On the one hand, it apparently requires the absolute probabilities per absolute unit time of all substitutions. In practice, however, that is not the case and only the relative probabilities are needed. If the $\pi_{ij}$ values are replaced by $k\pi_{ij}$ values and correspondingly fewer time units $t/k$ are used, then equation (6) remains identical and at an extremely close computational accuracy, for small $\pi_{ij}$ values:

$$M^t = \begin{pmatrix} (1 - k\Pi_1) & k\pi_{21} & \cdots \\ k\pi_{12} & (1 - k\Pi_2) & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \frac{t}{k}.$$  

On the other hand, even the relative probabilities of substitution cannot be estimated except from a very limited set of data. They cannot be estimated in general from the observed substitution between species, because the observed difference estimates the entries in the higher-order matrix $M^t$, not in the one-step substitution matrix $M$. To estimate relative one-step probabilities we need sequences that are sufficiently close together evolutionarily that multiple hits at a given position are rare. In practice this means that the mean number of hits per site must be less than, say, 0.25. Even for this low divergence, 12% of sites that have been hit will be multiple hits. But with so little divergence, most sites will not have changed at all, so that the number of sites that give information about the relative probabilities of different kinds of substitution will be quite small. In the absence of either a large data base on polymorphic sequences
within a species (Kreitman 1983) or comparisons between very closely related species (Schaeffer and Aquadro 1987), we simply have no reasonable estimates of the relative \( \pi_{ij} \)'s.

It turns out, however, that even this problem is not serious and that a close approximation to the process can be made with minimal information. We observe the following relations:

1. The Jukes-Cantor-Rice equation (2) is of the general form

\[
d = -b \ln \left( 1 - \frac{P}{b} \right),
\]

which can be rewritten as

\[
(b - P) = (b - P_0)e^{-d/b} \quad \text{when } P_0 = 0.
\]

This is the simplest description of the exponential decay of a difference between a variable \( P \) and its asymptotic value \( b \) as the number of events, \( d \), grows larger.

2. Tajima and Nei (1984), using a differential equation approach, studied a special case of the substitution matrix in which it is assumed that the total rate of substitution to a given base is the same for all bases when averaged over all states from which they come (the equal input model). The result to a very close approximation numerically is, in fact,

\[
d = -b \ln \left( 1 - \frac{P}{b} \right),
\]

where

\[
b = 1 - \sum g_i^2,
\]

the \( g_i \) being the average nucleotide frequencies in the compared segments.

3. Irrespective of the entries in the matrix \( M \), after a long period of evolution \( g' \) approaches an equilibrium vector, the eigenvector, with elements \( e_i \). These will be equal numerically to the elements on the principal diagonal of \( M' \) for large \( t \). Then, from equation (5),

\[
P = \sum_{i=1}^{4} g_i(1 - e_i) = 1 - \sum g_i e_i.
\]

But it is reasonable that the average nucleotide composition over the whole sequence, for both sequences, is close to the equilibrium reflecting the general probabilities of substitution, \( \pi_0 \). Then \( g_i \cong e_i \) and

\[
P \cong 1 - \sum e_i^2.
\]

Thus, if the general form (7) were a good approximation to the exact matrix multiplication method, \( 1 - \sum e_i^2 \) would be the appropriate value for \( b \) in the equation.

4. Extensive numerical calculations show that the eigenvector for the unit ei-
genvalue of the matrix $M$ is very sensitive to variations in the stabilities, $1 - \Pi_i$, but insensitive to variations in the off-diagonal elements.

All these point to the use of the relation (7) with $b = 1 - \sum e_i^2$ as a general approximation. We now note that, if there are $n$ possible states, then, by the definition of a variance and the fact that the mean of probabilities that add to unity is exactly $\mu = 1/n$, we obtain

$$\sum_{i=1}^{n} e_i^2 = n\sigma_e^2 + 1/n ;$$

$$d = -\frac{(n^* - 1)}{n^*} \ln \left( 1 - \frac{n^*}{n^* - 1} p \right),$$

where

$$n^* = \frac{n}{1 + \sigma^2_e} .$$

We can regard $n^*$ as an "effective" number of alternatives for a site. When $\sigma^2_e = 0$, then $n^* = n$, the actual number of alternatives. In what follows, I will call formula (8) the generalized Jukes and Cantor (GJC) formula.

The eigenvector, $e$, is the long-term equilibrium toward which the nucleotide frequencies will tend by evolutionary substitution as long as the elements of the substitution matrix remain constant. If we assume that the forces governing these base substitutions have remained more or less constant over a period that is long relative to the time of divergence of the two sequences being considered, then these sequences can themselves be regarded as samples from the equilibrium distribution. Then the frequencies of A, G, C, and T in these sequences provide an estimate of the $e_i$. This is the assumption made in the method of Tajima and Nei (1984).

To recapitulate, if it can be assumed that all the nucleotides in a sequence, irrespective of their position in the sequence, have the same set of probabilities of substitution, $\pi_{ij}$, and if sequences are close to equilibrium in their A, G, C, and T composition, then the GJC estimate gives a numerically accurate solution to the problem of inferring the number of evolutionary substitution events that have occurred between two sequences.

The question is, to what sequences does this assumption apply? In particular, in what sequences do all base positions have the same evolutionary dynamics? Since we have no evidence at present that mutation and repair processes differ from one base position to another (although differing between different bases), this question comes down to asking for what sequences are all nucleotide positions under the same functional constraints in evolution. Note that we do not require either that the nucleotides are unconstrained or that all four nucleotides are equally constrained but only that constraints are independent of serial position.

In the absence of evidence to the contrary, we assume that intergenic regions which neither code for protein nor regulate transcription qualify. Intron sequences, excluding the donor and acceptor ends, also seem to qualify. But coding sequences do not. Except for a protein that is totally unconstrained evolutionarily in its amino acid sequence so that all nucleotides are equally free to change to all other nucleotides, we
must distinguish synonymous from nonsynonymous changes. Nor can the problem for coding sequences be solved simply by analyzing first, second, and third positions of codons separately, because third positions, for example, are differently constrained depending on whether they are in a two-, three-, four-, or sixfold-redundant codon. Moreover, for twice-redundant codons, particular bases have different redundancies depending on whether the codon is T/C redundant or A/G redundant. So, even within first and third positions, all A's, say, are not equivalent in their evolutionary dynamics. Thus, the Tajima and Nei (1984) and Nei (1987) globin gene divergence analysis using the GJC method is inappropriate. On the other hand, their analysis of pseudogene divergence would be correct if all the divergence had occurred since the inactivation of the gene.

Analyzing Coding Sequences

In fact, no new theoretical derivation is needed to analyze coding sequences. It is simply a matter of the correct application of the GJC formula. For coding sequences, the codon—not the nucleotide—must be the unit of analysis. In what follows we assume that synonymous substitutions are much more probable than amino acid replacement substitutions. If that assumption is incorrect, an appropriate analysis is much more problematic (see below).

The assumption that nonsynonymous changes are relatively rare allows us a simple solution to the path problem, which is a necessary prerequisite to solving the multiplicity problem. If two compared sequences are identical for some codon, say, GAA (Glu), we need to estimate the probability that they have actually diverged and then converged. But that probability will depend on the intermediate states through which they have gone. If we can assume that only synonymous changes have occurred—or at least that nonsynonymous changes have a low probability—then the only two allowable states for this codon are GAA and GAG, since the Glu codon group is only twice redundant. Otherwise, we would have to consider all intermediate paths, for which we have no probability estimates.

We define a codon group as the set of codons that code for the same amino acid and are mutually accessible by synonymous base substitutions. Under this definition, serine is coded by two distinct codon groups, (AGT and AGC) and (TCT, TCC, TCA, and TCG), since the two groups are not accessible to each other only by synonymous substitutions. This separation is necessary in order that the substitution matrices be regular and thus converge to an equilibrium. We define a multiplicity class as the set of all codon groups that have the same redundancy number. We need concern ourselves only with multiplicity classes 2, 3, 4, and 6, since the once redundant codon groups Met and Tyr contain no information on synonymous substitutions.

For any individual codon group there is a substitution matrix analogous to matrix (4) but of a different size depending on the multiplicity class to which it belongs. For example, the matrix for Arg is given in table 1. The 0 entries in the matrix arise because no single-step substitution will carry, say, CGT into AGA, unlike the situation in the matrices for the 2, 3 and 4 multiplicity classes, where all states are accessible by a single step from all others. The substitution matrix for Leu would look the same as that for Arg, but, for example, the probability of CGA to AGA in Arg is in the same place as the probability of CTA to TTA in Leu although the two probabilities are not equal. While both are first-base-position changes, C to A is a transversion while C to T is a transition. Moreover, even when the base change is the same, as are CGT to CGG in Arg and CTT to CTG in Leu, there is (a) no reason to suppose that these
have the same probability in general and (b) strong reason, on the basis of known
codon biases, to suppose the contrary. (See the discussion of codon biases below.)

For any given codon group the number of evolutionary events per codon can be
estimated by the GJC formula (8), where $n^*$ is calculated from formula (9) as before:

$$n^* = \frac{n}{1 + n^2 \sigma_e^2},$$

where, again, $n = \text{multiplicity class of the codon group (2, 3, 4, or 6)}$ and $\sigma_e^2 = \text{the variance of the estimated equilibrium proportions of the different alternative synonymous codons in the codon group.}$

**Practical Estimation**

In principle, $d$ could be estimated separately for each of the 18 codon groups in
classes 2, 3, 4, and 6 and then averaged, since there is no reason to suppose that the
synonymous divergence rates will differ from group to group. In practice, however,
t here is too little information in any given sequence comparison to estimate each $d$
separately with any reliability. Given the unequal representation of different amino
acids, some codon groups will be represented only a few times. This leads to the
possibility of an infinite (or undefined) estimate of a given $d$, since, by chance, $[(n^* - 1)/n]P$ in formula (8) often may be greater than unity and the variance of $d$
estimates will be extremely large. The solution is to obtain a single $d$ estimate for each
multiplicity class by calculating a $\sigma_e^2$ for each codon group within the multiplicity
class and using the average, $\overline{\sigma_e^2}$, for a multiplicity class in relation (9) to obtain $n^*$
for that class. This value of $n^*$, substituted in formula (8), estimates $d$ for that class.
As we will see, these estimates for the separate multiplicity classes cannot be combined,
in general, into a single average value, because different classes evolve at differ-
ent rates.

Confidence limits for $d$ can be found directly by first finding the usual confidence
limits for the observed divergence $P$

$$P \pm t \sqrt{\frac{P(1 - P)}{N_c}},$$

where $N_c$ is the number of codons of multiplicity class $C$ in the sequence, and by then
substituting the upper and lower confidence limits for $P$ in formula (8). Calculation
Table 2
Codon usages for Xdh in *Drosophila melanogaster* (Keith et al. 1987; Lee et al. 1987) and *D. pseudoobscura* (Riley 1988) and Their Weighted Means Compared with the Average of 20 Peptides in *D. melanogaster* (Taken from the DNA Analysis Program of Pustell)

<table>
<thead>
<tr>
<th></th>
<th><em>D. melanogaster</em></th>
<th><em>D. pseudoobscura</em></th>
<th>Mean 20 Peptides</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phe</td>
<td>TTC 0.667</td>
<td>0.673</td>
<td>0.669</td>
</tr>
<tr>
<td></td>
<td>TTT 0.333</td>
<td>0.327</td>
<td>0.331</td>
</tr>
<tr>
<td>Tyr</td>
<td>TAC 0.561</td>
<td>0.786</td>
<td>0.675</td>
</tr>
<tr>
<td></td>
<td>TAT 0.439</td>
<td>0.214</td>
<td>0.325</td>
</tr>
<tr>
<td>His</td>
<td>CAC 0.667</td>
<td>0.634</td>
<td>0.649</td>
</tr>
<tr>
<td></td>
<td>CAT 0.333</td>
<td>0.366</td>
<td>0.351</td>
</tr>
<tr>
<td>Gln</td>
<td>CAG 0.761</td>
<td>0.893</td>
<td>0.826</td>
</tr>
<tr>
<td></td>
<td>CAA 0.239</td>
<td>0.107</td>
<td>0.174</td>
</tr>
<tr>
<td>Asn</td>
<td>AAC 0.512</td>
<td>0.600</td>
<td>0.557</td>
</tr>
<tr>
<td></td>
<td>AAT 0.488</td>
<td>0.400</td>
<td>0.443</td>
</tr>
<tr>
<td>Lys</td>
<td>AAG 0.762</td>
<td>0.882</td>
<td>0.816</td>
</tr>
<tr>
<td></td>
<td>AAA 0.238</td>
<td>0.118</td>
<td>0.184</td>
</tr>
<tr>
<td>Asp</td>
<td>GAC 0.350</td>
<td>0.638</td>
<td>0.513</td>
</tr>
<tr>
<td></td>
<td>GAT 0.650</td>
<td>0.362</td>
<td>0.487</td>
</tr>
<tr>
<td>Glu</td>
<td>GAG 0.731</td>
<td>0.902</td>
<td>0.818</td>
</tr>
<tr>
<td></td>
<td>GAA 0.269</td>
<td>0.097</td>
<td>0.182</td>
</tr>
<tr>
<td>Cys</td>
<td>TGC 0.718</td>
<td>0.743</td>
<td>0.722</td>
</tr>
<tr>
<td></td>
<td>TGT 0.282</td>
<td>0.257</td>
<td>0.278</td>
</tr>
<tr>
<td>Ser</td>
<td>AGC 0.536</td>
<td>0.781</td>
<td>0.667</td>
</tr>
<tr>
<td></td>
<td>AGT 0.464</td>
<td>0.219</td>
<td>0.333</td>
</tr>
<tr>
<td>Ile</td>
<td>ATC 0.514</td>
<td>0.542</td>
<td>0.528</td>
</tr>
<tr>
<td></td>
<td>ATT 0.357</td>
<td>0.347</td>
<td>0.352</td>
</tr>
<tr>
<td></td>
<td>ATA 0.129</td>
<td>0.111</td>
<td>0.120</td>
</tr>
<tr>
<td>Val</td>
<td>GTG 0.604</td>
<td>0.602</td>
<td>0.603</td>
</tr>
<tr>
<td></td>
<td>GTC 0.156</td>
<td>0.265</td>
<td>0.211</td>
</tr>
<tr>
<td></td>
<td>GTT 0.167</td>
<td>0.031</td>
<td>0.098</td>
</tr>
<tr>
<td></td>
<td>GTA 0.073</td>
<td>0.102</td>
<td>0.088</td>
</tr>
<tr>
<td>Ser</td>
<td>TCC 0.392</td>
<td>0.404</td>
<td>0.398</td>
</tr>
<tr>
<td></td>
<td>TCG 0.353</td>
<td>0.381</td>
<td>0.366</td>
</tr>
<tr>
<td></td>
<td>TCA 0.137</td>
<td>0.143</td>
<td>0.140</td>
</tr>
<tr>
<td></td>
<td>TCT 0.118</td>
<td>0.072</td>
<td>0.096</td>
</tr>
<tr>
<td>Pro</td>
<td>CCC 0.408</td>
<td>0.452</td>
<td>0.431</td>
</tr>
<tr>
<td></td>
<td>CCG 0.225</td>
<td>0.370</td>
<td>0.299</td>
</tr>
<tr>
<td></td>
<td>CCT 0.169</td>
<td>0.137</td>
<td>0.152</td>
</tr>
<tr>
<td></td>
<td>CCA 0.197</td>
<td>0.041</td>
<td>0.118</td>
</tr>
</tbody>
</table>
Table 2 (Continued)

<table>
<thead>
<tr>
<th></th>
<th>D. melanogaster</th>
<th>D. pseudoobscura</th>
<th>Mean</th>
<th>20 Peptides</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thr:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACC</td>
<td>0.425</td>
<td>0.385</td>
<td>0.404</td>
<td>0.633</td>
</tr>
<tr>
<td>ACG</td>
<td>0.260</td>
<td>0.346</td>
<td>0.305</td>
<td>0.257</td>
</tr>
<tr>
<td>ACT</td>
<td>0.178</td>
<td>0.141</td>
<td>0.159</td>
<td>0.096</td>
</tr>
<tr>
<td>ACA</td>
<td>0.137</td>
<td>0.128</td>
<td>0.132</td>
<td>0.016</td>
</tr>
<tr>
<td>Ala:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GCC</td>
<td>0.466</td>
<td>0.532</td>
<td>0.500</td>
<td>0.635</td>
</tr>
<tr>
<td>GCT</td>
<td>0.220</td>
<td>0.210</td>
<td>0.215</td>
<td>0.184</td>
</tr>
<tr>
<td>GCG</td>
<td>0.178</td>
<td>0.153</td>
<td>0.165</td>
<td>0.095</td>
</tr>
<tr>
<td>GCA</td>
<td>0.138</td>
<td>0.105</td>
<td>0.120</td>
<td>0.086</td>
</tr>
<tr>
<td>Gly:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GGC</td>
<td>0.321</td>
<td>0.578</td>
<td>0.450</td>
<td>0.500</td>
</tr>
<tr>
<td>GGA</td>
<td>0.376</td>
<td>0.229</td>
<td>0.302</td>
<td>0.270</td>
</tr>
<tr>
<td>GGT</td>
<td>0.220</td>
<td>0.119</td>
<td>0.170</td>
<td>0.214</td>
</tr>
<tr>
<td>GGG</td>
<td>0.083</td>
<td>0.074</td>
<td>0.078</td>
<td>0.016</td>
</tr>
<tr>
<td>Leu:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CTG</td>
<td>0.492</td>
<td>0.557</td>
<td>0.523</td>
<td>0.571</td>
</tr>
<tr>
<td>CTC</td>
<td>0.161</td>
<td>0.270</td>
<td>0.215</td>
<td>0.114</td>
</tr>
<tr>
<td>TTG</td>
<td>0.186</td>
<td>0.070</td>
<td>0.129</td>
<td>0.114</td>
</tr>
<tr>
<td>CTG</td>
<td>0.102</td>
<td>0.052</td>
<td>0.077</td>
<td>0.118</td>
</tr>
<tr>
<td>CTA</td>
<td>0.042</td>
<td>0.043</td>
<td>0.043</td>
<td>0.036</td>
</tr>
<tr>
<td>TTA</td>
<td>0.017</td>
<td>0.008</td>
<td>0.013</td>
<td>0.046</td>
</tr>
<tr>
<td>Arg:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CCG</td>
<td>0.266</td>
<td>0.462</td>
<td>0.368</td>
<td>0.458</td>
</tr>
<tr>
<td>CCG</td>
<td>0.200</td>
<td>0.200</td>
<td>0.200</td>
<td>0.073</td>
</tr>
<tr>
<td>CGT</td>
<td>0.160</td>
<td>0.138</td>
<td>0.148</td>
<td>0.207</td>
</tr>
<tr>
<td>CGA</td>
<td>0.147</td>
<td>0.125</td>
<td>0.135</td>
<td>0.106</td>
</tr>
<tr>
<td>AGA</td>
<td>0.147</td>
<td>0.012</td>
<td>0.077</td>
<td>0.045</td>
</tr>
<tr>
<td>AGG</td>
<td>0.080</td>
<td>0.063</td>
<td>0.071</td>
<td>0.112</td>
</tr>
</tbody>
</table>

of these limits will be a sobering experience because small confidence limits in $P$ may be converted to very large limits on $d$, especially near saturation.

**Estimation of $e_i$**

The question arises as to the best estimate of the equilibrium frequencies. We might assume that the observed codon usages represent the equilibria (the eigenvectors) of the substitution matrices for each codon group. This assumption is supported by the relative stability of codon usages across related species, despite extensive synonymous substitution. For example, Schaeffer and Aquadro (1987) found no significant difference in codon usage for Adh between *Drosophila pseudoobscura* and *D. melanogaster*, despite a 33% synonymous divergence in their sequences. Even when there is a significant difference between codon usages, as between the Xdh sequences of *D. pseudoobscura* and *D. melanogaster* (Riley 1988), the order of use of the bases for each codon group is generally the same in the two species. But the similarity of codon usage of a gene in several related species is not a convincing argument that this usage is at equilibrium. It is equally plausible that it is a similarity resulting from common origin and is a historical relic of the usage in the common ancestor.

If we choose to use known sequences themselves to estimate the $e_i$, then there are several possibilities. An average of the two compared sequences may be used—or
Table 3
Estimates of $\sigma^2$, $n^*$, and $d$ between Xdh Sequences of Drosophila melanogaster and D. pseudoobscura, Based on Four Different Estimates of Equilibrium $e_i$

<table>
<thead>
<tr>
<th>Multiplicity Class and $P$</th>
<th>$D$. melanogaster Xdh</th>
<th>$D$. pseudoobscura Xdh</th>
<th>Average Xdh</th>
<th>Average $D$. melanogaster</th>
<th>Equal Input/Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>2:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>0.0322</td>
<td>0.0759</td>
<td>0.0468</td>
<td>0.0752</td>
<td>0</td>
</tr>
<tr>
<td>$n^*$</td>
<td>1.772</td>
<td>1.534</td>
<td>1.685</td>
<td>1.538</td>
<td>2</td>
</tr>
<tr>
<td>$d$</td>
<td>0.72</td>
<td>Above saturation</td>
<td>0.81</td>
<td>Above saturation</td>
<td>0.60</td>
</tr>
<tr>
<td>3:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>0.0250</td>
<td>0.0310</td>
<td>0.0279</td>
<td>0.1102</td>
<td>0</td>
</tr>
<tr>
<td>$n^*$</td>
<td>2.449</td>
<td>2.346</td>
<td>2.403</td>
<td>1.506</td>
<td>3</td>
</tr>
<tr>
<td>$d$</td>
<td>0.77</td>
<td>0.89</td>
<td>0.78</td>
<td>Above saturation</td>
<td>0.69</td>
</tr>
<tr>
<td>4:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>0.0180</td>
<td>0.0297</td>
<td>0.0218</td>
<td>0.0400</td>
<td>0</td>
</tr>
<tr>
<td>$n^*$</td>
<td>3.106</td>
<td>2.712</td>
<td>2.966</td>
<td>2.438</td>
<td>4</td>
</tr>
<tr>
<td>$d$</td>
<td>1.13</td>
<td>1.30</td>
<td>1.18</td>
<td>1.59</td>
<td>0.99</td>
</tr>
<tr>
<td>6:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma^2$</td>
<td>0.0140</td>
<td>0.0293</td>
<td>0.0198</td>
<td>0.0266</td>
<td>0</td>
</tr>
<tr>
<td>$n^*$</td>
<td>3.990</td>
<td>2.920</td>
<td>3.503</td>
<td>3.063</td>
<td>6</td>
</tr>
<tr>
<td>$d$</td>
<td>2.23</td>
<td>Above saturation</td>
<td>3.82</td>
<td>Above saturation</td>
<td>1.47</td>
</tr>
</tbody>
</table>

an average over a much larger sample of sequences, say, the average of all coding sequences in the species or the average of all species for the particular coding sequence. When codon usages are quite unequal, the variances, $\sigma^2$, and thus the estimates of amount of evolutionary divergence are quite sensitive to small differences in codon usage. An example of the variability of estimates with different codon usage schedules can be seen in table 3. The table shows the values of $\sigma^2$, $n^*$, and estimated divergence for the Xdh sequences of D. melanogaster and D. pseudoobscura (see Riley 1988). Four different sets of estimates of $\sigma^2$ are compared: one from the D. melanogaster sequence, one from the D. pseudoobscura sequence, one from the average of the two sequences, and one from the average of 20 protein coding sequences in D. melanogaster. Clearly the estimates of $e_i$ from an average over many different genes is unsatisfactory. In both D. melanogaster and D. pseudoobscura the degree of codon bias for Xdh is consistently less than the bias for the average gene. For example, in D. melanogaster every codon group but one has a more even use of alternative codons than is the case for the average of all genes (see table 2). So, the variances of the $e_i$ for Xdh are less than those for an “average” gene. The divergence that has occurred for Xdh is larger than the saturation level predicted from the “average” gene, and so we get infinite estimates of $d$. The codon usage of Xdh itself is a more appropriate estimation of the equilibrium $e_i$ for Xdh. But even a single species will give some impossible estimates, as the results for D. pseudoobscura Xdh show. It seems clear that an average over several closely related species for a specific gene is needed to get adequate estimates of the $g_i$. When the average divergence is low, $<0.25$, this instability is not serious, since the divergence rate is insensitive to redundancy below this level. At the level of divergence shown by Xdh, however, this instability can be a serious impediment to accurate estimation. The source of the instability of the estimates of $d$ can be seen in figure 1. It shows the relation between observed sequence difference, $P$, and estimated $d$ for different values of $n^*$. When the observed difference is $> \sim 0.4$, it makes an
immense difference to the estimate of \( d \) which curve is used to project \( d \). A small sampling variation in either \( n^* \) or \( P \) can result in an infinite \( d \), i.e., one in which the asymptote is actually below the observed \( P \). An important lesson to be drawn from table 3 and figure 1 is that estimates of evolutionary divergence are to be taken with liberal doses of salt when observed divergences, \( P \), are near saturation.

The sensitivity of the estimates of divergence to the sequences used to estimate the \( e_i \) has both advantages and disadvantages. On the positive side, if the sequences are indeed near their equilibrium, and if different genes or different taxa have different equilibria, the comparisons of divergences of different genes or of different taxa must take these codon usage differences into account. On the negative side, if differences in codon usage between genes and taxa are not equilibrial but historical differences, the estimation of divergences will show a false variation among genes and taxa that is a consequence of the mistaken equilibrium estimates.

The alternative to using actual codon usages is to make model assumptions about the actual process of substitution and to calculate the theoretical eigenvector. In the absence of detailed knowledge of \( \pi_{ij} \) in the substitution matrix, one might make various simplifying assumptions. If it is assumed, for example, that the total rate of substitution to a codon from all alternatives within the group is equal to the total rate of substitution away from that codon, then the matrix is doubly stochastic, from which it follows that the eigenvector is simply \((1/n, 1/n, 1/n, \ldots)\) and \( \sigma^2 = 0 \) irrespective of the \( \pi_{ij} \). The result of this equal input/output model is shown in the last column of table 3.

There are advantages and disadvantages to this method of estimating the \( e_i \). On the positive side, the wide fluctuations and instability of the estimates of \( d \) that arise from using a restricted base of codon usages is avoided. It is less likely that artifactual differences in estimated \( d \) will appear from historical or sampling differences in observed codon usage. Estimates of \( d \) on this model are always conservative. On the negative side, the \( e_i \) are unlikely to be really equal and some apparent difference in divergence rates between different taxa or different genes may be really the result of different equilibrium \( e_i \)—and may not truly represent different rates of divergence. A special case of this model is one in which the stabilities \((1 - \Pi_i)\) of all the codons within the group are assumed equal. It then follows that all of the off-diagonal elements become equal, and we return to the original isotropic model for which the Cantor-Jukes formula was originally derived.

An alternative model makes the assumption that the stabilities of the codons and the rates of change to them are arbitrary but that, if a codon changes, it is equally likely to change to any of the alternatives. So, the \( \Pi_i \) are unequal but

\[ \pi_j = \frac{\Pi_i}{(n - 1)} \text{ for all } j \neq i. \]

In this case it is easy to verify by multiplication that the eigenvector corresponding to the unit eigenvalue has the components

\[ g_i = \frac{1}{\Pi_i} = \frac{1}{\Pi_i} \frac{H(\Pi)}{n}, \]

where \( H(\Pi) \) is the harmonic mean of the \( \Pi_i \). This model is useful for two reasons.
First, numerical calculations show that the eigenvector of the matrix $M$ is much more sensitive to variations in the $\Pi_i$ than to variation in the off-diagonal $\pi_{ij}$ for a given $\Pi_i$. Thus, even though there will be variation in the $\pi_{ij}$, the result will depend chiefly on estimates of the stabilities $(1 - \Pi_i)$. Second, estimates of stability require only the frequency at which a given codon changes, pooled over all the alternatives to which it changes, so a smaller data base is needed to estimate the relative $\Pi_i$ than is needed for all the $\pi_{ij}$ ($\sim$40% as much data averaged over multiplicity classes).

Finally, we may assume that the stabilities are equal but that there is some predetermined pattern of substitution from a given codon. A particular case is to allow for a transition/transversion bias. This equal stability model has the advantage that the absolute stabilities are irrelevant because an exact calculation of divergence can be made by a slightly different route, given in the Appendix. If the model used does, in fact, assume that the only differences in substitution rates are a consequence of transition/transversion bias, as shown in the Appendix, the equal stability model is essentially indistinguishable from the isotropic model. Riley (1988) employed the equal stability model but used the exact calculation procedure given in the Appendix, rather than using the GCJ. Her results differ only slightly from those given in the last column of table 2.

**Nucleotides as Units of Analysis**

Table 4, in conjunction with table 3, shows us why we cannot use averages over single base positions, even when first, second, and third positions are treated separately.
Table 4
Nucleotide Frequencies for Xdh Coding Region, Average over *Drosophila melanogaster* and *D. pseudoobscura*, Together with Divergence Calculated between Species on a Nucleotide Basis

<table>
<thead>
<tr>
<th>POSITION</th>
<th>First</th>
<th>Second</th>
<th>Third</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>0.1510</td>
<td>0.2852</td>
<td>0.1795</td>
<td>0.2053</td>
</tr>
<tr>
<td>C</td>
<td>0.2422</td>
<td>0.2361</td>
<td>0.3703</td>
<td>0.2830</td>
</tr>
<tr>
<td>A</td>
<td>0.2475</td>
<td>0.2837</td>
<td>0.1113</td>
<td>0.2142</td>
</tr>
<tr>
<td>G</td>
<td>0.3593</td>
<td>0.1949</td>
<td>0.3383</td>
<td>0.2977</td>
</tr>
<tr>
<td>$\sigma_g^2$</td>
<td>0.005454</td>
<td>0.001401</td>
<td>0.01164</td>
<td>0.001661</td>
</tr>
<tr>
<td>&quot;n*&quot;</td>
<td>3.6789</td>
<td>3.9123</td>
<td>3.3720</td>
<td>3.8964</td>
</tr>
<tr>
<td>&quot;q&quot;</td>
<td>0.202</td>
<td>0.202</td>
<td>0.204</td>
<td>0.203</td>
</tr>
</tbody>
</table>

**NOTE.—**Observed nucleotide divergence = 0.177.

Table 4 gives the proportions of bases, by position in the codons of Xdh, computed from the same sequences used to calculate table 3. When calculated on pooled bases, the variances, $\sigma_g^2$, are smaller than the average variances shown in table 3. A fictitious "n*" calculated from the variances in table 4 is then too large for the GJC formula. Moreover, the distributions in table 4 arise from quite different processes. For second-position bases, the distributions are purely a consequence of the amino acid composition of the proteins and so reflect functional and selective constraints of unknown form. First-position base frequencies are a mixture of protein constraints and synonymous substitution processes. The contributions of multiplicity class 2, 3, and 4 codons to this distribution arise entirely from amino acid constraints, but class 6 contributes to this distribution both from synonymous processes and amino acid constraints. Even the third-position bases have a distribution that is a mixture of synonymous and amino acid constraints. The contributions to this distribution from class 2 and class 6 codons reflect both the amino acid distribution and the synonymous substitution processes, and these codons account for only 39% of the third-position bases. Thus, an analysis by base positions used, for example, by Tajima and Nei (1984), confounds the constraints on amino acid substitutions with the synonymous substitution dynamics. An analysis using nucleotides as the units of analysis is valid only when there are no significant amino acid constraints. Several methods have been developed that recognize codon differences in the computation of average nucleotide divergence but that carry out the divergence calculation on a nucleotide basis. These include the methods of Miyata and Yasunaga (1981), Nei and Gojobari (1986), and several others reported by Nei (1987). These techniques classify nucleotide positions as synonymous or nonsynonymous codon by codon. Moreover, they distinguish between two-, three-, and fourfold synonymous sites, depending on how many possible synonymous changes can take place at the site. In addition, for codon differences that involve more than one observed base difference, they reconstruct all the shortest pathways connecting the two forms, assign each path a probability (by a rule that differs from author to author), and add the inferred changes to the directly observed cases of synonymous and nonsynonymous differences. Having now calculated a proportion of substituted positions separately for synonymous positions, $P_s$, and for nonsynonymous positions, $P_n$, one can substitute these values separately in the Jukes-Cantor
formula to derive separate estimates of \( d \) for the two classes. But this last step is invalid for two reasons. First, the standard Jukes-Cantor formula applies only to the fourfold multiple nucleotide positions. No account is taken of the variance, \( \sigma^2 \), so it is not clear how an effective \( n^* \) is to be computed in order to use the GJC formula instead.

The second problem with the method is deeper. The nonsynonymous differences cannot be treated as a multihit process in the same manner as the synonymous changes. Suppose we observe a single nonsynonymous difference, say, ACT (Thr) to GCT (Ala). The multiplicity problem is that \( 1, 2, 3, \ldots, n \) evolutionary substitutions may have occurred between these two codons. But what kind of substitutions (the path problem)? Surely, we cannot assume that all the substitutions were nonsynonymous. Most of the steps in the multistage process that may have taken place between ACT and GCT belong to the synonymous class. There is thus a contradiction between assuming that synonymous and nonsynonymous changes belong to different probability classes, on the one hand, and using a formula that assumes a homogeneous process, on the other. In general, unless amino acid substitutions are unconstrained, no information about synonymous changes can be derived reliably from amino acid substitutions. A correct analysis must use data only from codons that are synonymous between the compared sequences.

**Amino Acid Constraints**

If amino acids are almost completely unconstrained in their substitutions, then the analysis by redundancy classes is incorrect. Every nucleotide has four possible states, and the correct analysis is given by Tajima and Nei (1984). They give a slightly improved version of equations (8) and (9) that makes optimal use of the information from the two compared sequences to provide an estimate of the equilibrium vector \( \epsilon \) but that is then tied to actual variation between those sequences.

If amino acids are very highly constrained, then the analysis described here by codon redundancy groups is the correct one. As a rule of thumb, if synonymous changes are 10 times as likely as amino acid changes, the codon analysis will not be far off, since the probability of a mixed synonymous/nonsynonymous pathway of given length will be of a lower order of magnitude than the probability of purely synonymous pathways.

Evidence about the relative probability of synonymous and nonsynonymous substitutions cannot be derived from comparisons of distantly related species. Even when replacement substitutions have a low probability relative to synonymous substitutions, as time goes on the ratio of observed amino acid changes to observed synonymous changes increases. The synonymous changes will saturate relatively quickly, approaching some limit given by the substitution probability matrix. Amino acid changes will occur more slowly, but after a long time these too will approach a limit, set by the functional constraints of the protein. The actual observed ratio of replacement to synonymous changes then will reflect more and more the relative equilibrium values rather than the per-time-unit substitution probabilities. The evidence from studies of polymorphism within species (Kreitman 1983) and between very closely related species (Schaeffer and Aquadro 1987; Riley 1988) for the Adh and Xdh genes in *Drosophila* shows that amino acid replacements are indeed very much less likely than silent changes. There can be exceptions, however, such as the chorion protein genes of *Drosophila*, which, at very low divergences, show that amino acid replacements are about half as probable as silent substitutions (Martinez-Cruzado 1988).

It makes a huge difference to the result whether the estimate is made on a nu-
cleotide or codon basis, as a comparison between tables 3 and 4 shows. Despite the variability in the codon-based result, depending on the codon usages that are employed as an estimate of the equilibrium $e$, the nucleotide-based estimate is well below all the feasible estimates based on codons.

The question arises, finally, about what is to be done with proteins that fit neither model, having a significant but not extreme selection against amino acid replacements—say, a three or four times greater probability of a synonymous substitution. In principle, nothing new is needed. The complete set of 61 sense codons will be connected to each other by a $61 \times 61$ substitution matrix $M$. This one-step matrix will look like table 1, with zero entries for every substitution that is not possible in a single step. The second power, $M^2$, will continue to show some zero entries, since two steps are not sufficient to connect all codons; but by $M^3$ all entries are nonzero, and the eigenvectors of the original matrix are the predicted equilibrium. But, in fact, no equilibrium will be approached because the matrix $M$ is not a constant. Precisely because there are functional and selective constraints on the probability of a given nonsynonymous change, as soon as that substitution has occurred the substitution probabilities will change in some unknown manner. That is, the selective constraints on amino acid sequence change as the sequence changes! It is a contradiction to claim, on the one hand, that amino acid substitutions are not neutral and then, on the other, to use a constant probability substitution matrix. There is in principle no way to reconstruct the number of evolutionary events that have occurred if every path has an unrecoverable probability of occurrence.

Summary Protocol

To summarize the discussion, we may outline a method for estimating the amount of evolutionary divergence between two coding sequences (mean number of evolutionary events per codon or nucleotide).

I. From data comparing sequences with a low observed proportion of synonymous changes (ideally, $<20\%$), determine the relative probability of amino acid replacement changes. (For most empirical compositions we expect three times as many amino acid replacements as synonymous substitutions if all code substitutions are equally likely. So if we observe an equal number of synonymous and nonsynonymous substitution, the probability of amino acid replacement is only about one-third that of synonymous changes.) If the relative probability of an amino acid replacement is
   A. $<10\%$ of synonymous, go to III;
   B. $>75\%$ of synonymous, go to II;
   C. $10\%$–$75\%$ of synonymous, go to IV.

II. Amino acid replacements are virtually unconstrained. Use formulas (8) and (9), where the $e_i$ are the frequencies of the four nucleotides in the entire sequence. A somewhat better approximation is given by the method of Tajima and Nei (1984), but there is no use in separating out first-, second-, and third-position nucleotides. The result is a mean number of evolutionary substitutions per nucleotide.

III. Amino acid replacements are highly constrained. The codon is the appropriate unit of analysis.
   A. Only codons that are synonymous should be compared.
B. The codons should be separated into multiplicity classes 2, 3, 4, and 6 and treated separately. Serine codons should be divided into a multiplicity class 4 set and a multiplicity class 2 set and should be added to the rest of the data for those multiplicity classes.

C. Calculate the observed proportion of substituted codons, $P$, for each multiplicity class.

D. Choose a model for estimating the eigenvector $e_i$.
   1. Codon usage model:
      a. Choose a codon usage table, say, the average of the codon usages in the two sequences being compared.
      b. Calculate the variance of codon usage proportions, $\sigma^2$, for each codon group separately in each multiplicity class.
      c. Take the weighted average of the $\sigma^2$ over all codon groups within the multiplicity class, weighted by codon group frequency.
      d. Use the average $\sigma^2$ as an estimate of $\sigma^2$ in formulas (8) and (9), with $n = $ the multiplicity class number. The result is a mean number of substitutions per codon.
   2. Constant stability model, with codon bias: Use formula (8) with $n^* = $ the multiplicity class number for $n = 2, 3, 4$ and $n^* = 5.6$ for $n = 6$. The result is a mean number of substitutions per codon.

IV. Amino acids are significantly but not overwhelmingly constrained. Nothing. No estimates of divergence are trustworthy.

A computer program available from the author compares two aligned coding sequences and provides

1. A listing of all synonymous and nonsynonymous changes;
2. Inferred paths of multiple substitutions where more than one nucleotide is changed in a codon—where the most parsimonious path therefore is given, involving the fewest amino acid replacements;
3. codon usage tables for the sequences;
4. a summary of observed and estimated divergences by multiplicity class using either the codon usage model, the constant stability model, or an arbitrary stability model.

Acknowledgments

The work described here was supported by NIH grant 5R01GM21179-14. I am grateful for the substantial intellectual contributions of Hamish Spencer, Margaret Riley, and Stephen Schaeffer and for the clarifications requested by anonymous reviewers.

APPENDIX

Exact Calculation of Divergence for the Equal Stability Model

If all the stabilities are equal, then we may consider the matrix of conditional probabilities that state $i$ will change to state $j$ given that a substitution has occurred.
This matrix (for the case of four alternative states) looks as follows:

\[
C = \begin{bmatrix}
0 & P_{21} & P_{31} & P_{41} \\
P_{12} & 0 & P_{32} & P_{42} \\
P_{13} & P_{23} & 0 & P_{43} \\
P_{14} & P_{24} & P_{34} & 0
\end{bmatrix}
\]

\[\sum_j P_{ij} = 1 \text{ for all } i\]

The matrix has zeroes on the principal diagonal because a substitution has taken place. Then, the probability that a codon starting in state \(i\) will be found in state \(j\) after exactly \(k\) evolutionary events will be \(P_{ij}^{(k)}\), the \(i, j\)th element in the matrix \(C^k\), the \(k\)th power of matrix \(C\). The probability that a codon will be found in a state identical with its starting state will then be \(P_{ii}^{(k)}\).

Suppose that the average number of substitution events that have occurred in the sequence is \(d\). Some codons will not have been substituted, some will have changed once, twice, etc., and the probabilities of these numbers of changes will be given by a Poisson distribution with mean \(d\). Therefore, the observed proportion of substituted codons \(P\), will be given by

\[
P = 1 - \sum_{k=1}^{\infty} P_{ii}^{(k)} e^{-d} \frac{d^k}{k!}.
\]

In practice, one calculates the right-hand side of formula (10) for increasing values of \(d\) until it agrees with the value of \(P\). The \(P_{ii}^{(k)}\) need be calculated only once for \(k\) from 1 to 10 to get all the accuracy needed.

A reasonable choice for the \(P_{ij}\) in the matrix \(C\) is to suppose that they are equal except for a transition/transversion bias. If we denote \(P_s\) the one-step \(P_{ii}\) for a transition event and \(P_v\) the one-step \(P_{ij}\) for a transversion event and assign

\[
B = P_s / P_v,
\]

then the matrix \(C\) can be written, for each multiplicity class, entirely in terms of \(B\), with no other parameters, because, by hypothesis, there are only two numerically different \(P_{ij}\)'s, one for transitions and one for transversions, and the \(P_{ij}\)'s add to unity for each \(i\). Substitution for the \(P_{ij}\)'s in terms of \(B\) shows that the matrices for multiplicity classes 2, 3, and 4 are, in fact, double stochastic and so have eigenvectors \((1/n, 1/n, 1/n, \ldots)\), irrespective of the actual bias, \(B\). The matrices for the two class 6 amino acids, arginine and leucine, are not precisely doubly stochastic, but their eigenvectors are quite insensitive to a wide variation in \(B\), as the following calculations show.

\[
\begin{array}{ccc}
B & \sigma^2_i & n^* \text{ [from eq. (9)]} \\
0.5 & 0.002958 & 5.42 \\
1.0 & 0.001853 & 5.62 \\
10.0 & 0.001176 & 5.26 \\
\end{array}
\]

When transition/transversion ratios are very high, the rate of approach to the equilibrium vector is slow, so that a better estimation procedure would be to use the exact iterative procedure of equation (10) for multiplicity classes 3 and 4 (class 2 consists only of transitions), especially if the observed divergence is in the range where
the estimated divergence is \( \approx 1 \). This is the procedure followed by Riley (1988). For example, the Adh sequences of *Drosophila melanogaster* and *D. pseudoobscura* differ by an observed 48% silent substitutions (Schaeffer and Aquadro 1987). The estimated divergences between the sequences for different assumed transition/transversion biases \( (B) \) are as follows:

\[
\begin{array}{cc}
B & d \\
1.00 \text{ (eigenvector)} & 0.77 \\
1.75 \text{ (observed)} & 0.80 \\
10.00 & 0.95 \\
100.00 & 1.40 \\
\end{array}
\]

In fact, \( B \) values >2.50 have not been found for sequence comparisons from nature, so the error incurred from using the eigenvector is not great. However, the estimates of \( B \) from observed organisms are downwardly biased because they represent high-order substitution probabilities because of multiple hits, rather than the one-step substitution probabilities, \( P_{ij} \).

**LITERATURE CITED**


Rice, K. 1987. DIV and RSP algorithms, commands and outputs. Department of Computer Sciences, University of Georgia.


**BARRY G. HALL**, reviewing editor

Received March 7, 1988; revision received July 20, 1988