Molecular Evolution Modeled as a Fractal Poisson Process in Agreement with Mammalian Sequence Comparisons

David R. Bickel and Bruce J. West
Center for Nonlinear Science, University of North Texas, Denton

The fractal doubly stochastic Poisson process (FDSPP) model of molecular evolution, like other doubly stochastic Poisson models, agrees with the high estimates for the index of dispersion found from sequence comparisons. Unlike certain previous models, the FDSPP also predicts a positive geometric correlation between the index of dispersion and the mean number of substitutions. Such a relationship is statistically proven herein using comparisons between 49 mammalian genes. There is no characteristic rate associated with molecular evolution according to this model, but there is a scaling relationship in rates according to a fractal dimension of evolution. The FDSPP is a suitable replacement for the homogeneous Poisson process in tests of the lineage dependence of rates and in estimating confidence intervals for divergence times. As opposed to other fractal models, this model can be interpreted in terms of Darwinian selection and drift.

Introduction

Kimura’s (1983) neutral theory of molecular evolution in its simplest form requires the rate of molecular substitutions to be roughly constant in time, which is the molecular-clock hypothesis of Zuckerkandl and Pauling (1965). This stochastic clock is typically modeled by a homogeneous (constant rate) Poisson process, which has an index of dispersion (ratio of the variance to the mean number of substitution events) of unity, since the variance equals the mean. Kimura used star phylogenies of various species to estimate the index of dispersion of the number of substitutions that occurred in various species since their divergence from a common ancestor. Applying chi-square tests to various star phylogenies, Kimura argued that the indices of dispersion are not significantly greater than unity for most star phylogenies tested and thus concluded that the null hypothesis of the homogeneous Poisson process could not be rejected, in agreement with his theory of the neutrality of substitutions.

Goldman (1994) pointed out that when star phylogenies are used to test the Poisson molecular clock, an elevated index of dispersion can result from lineage-dependent effects rather than from a time variation in substitution rates. Possible lineage effects include different substitution rates in different lineages or different divergence times for different lineages. Goldman performed Monte Carlo simulations to determine the probability distributions of index of dispersion estimators both for star phylogenies, in which the divergence times are equal, and for other phylogenies, in which the divergence times are unequal. Using these distributions, he found that while the index of dispersion estimate for \( \alpha \)-hemoglobin was sufficiently high to reject the homogeneous Poisson clock assuming a star phylogeny, it was not sufficiently high to reject it without that assumption. He also showed that calculating similar Monte Carlo distributions for nonstar phylogenies did not compel the rejection of the clock for cytochrome oxidase 2 or for the \( \gamma \)-psuedogene.

In a more comprehensive study considering 20 genes, Gillespie (1991), also aware of the problem of assuming equal divergence times, analyzed three-species trees rather than star phylogenies in his estimations of the index of dispersion. He merely assumed that each of three mammalian species evolved from a common ancestor, instead of assuming that they all branched from that ancestor at roughly the same time. Gillespie removed lineage effects by including weighting factors in his estimators of the index of dispersion and found that the homogeneous Poisson protein clock could be rejected at the 5% level for 12 of the 20 genes considered. He concluded that since the elevation in the weighted index-of-dispersion estimate did not result from lineage-dependent rates or lineage-dependent branch times, it had to be the result of substitution rates that change in time.

Takahata (1991) considered another possible cause of high index-of-dispersion estimates, the effect of the rate variation between sites of a gene according to the common gamma distribution suggested by Uzzell and Corbin (1971). Combining the homogeneous Poisson process with the common gamma distribution of rates across sites, Takahata (1991) proved that the expected index of dispersion between lineages is unity assuming that each site in a gene evolves at the same rate in all lineages. He concluded that the variation of rates among sites is insufficient to account for elevated indices of dispersion, which he attributed to time-varying substitution rates.

Gillespie (1991) generalized the homogeneous Poisson clock model by allowing the rate of the process to change in time. This process is called doubly stochastic, since it has a random rate and a random number of events given a rate. The doubly stochastic Poisson process has an index of dispersion higher than unity and thus agrees with estimations of the index of dispersion from sequence data. Gillespie’s (1991) particular choices of rate distributions and limiting conditions predict an index of dispersion that is independent of the mean number of substitutions. However, this prediction disa-
degrees with Ohta’s (1995) recent index-of-dispersion estimates from a three-species tree using 49 genes, since those estimates increase with the estimated mean numbers of substitutions, as we demonstrate below.

We thus propose a fractal doubly stochastic Poisson process (FDSPP) model of molecular evolution as a natural extension to the nonfractal doubly stochastic Poisson process (NFDSPP) models used by Gillespie (1991). Like Gillespie’s models, the FDSPP model agrees with the estimated overdispersion of substitutions, but unlike his models, it also agrees with the strong correlation between the estimated indices of dispersion and means of the numbers of substitutions.

**Doubly Stochastic Poisson Process**

Consider a stationary rate \( \lambda(T) \) that is described by its mean \( \rho \) and its autocovariance \( \Gamma(\tau) \), given that \( T \) is an instantaneous point in time and \( \tau \) is a lag time between rates. A Poisson process corresponding to this rate has an intensity of

\[
\Lambda(t) = \int_0^t \lambda(T) \, dT \tag{1}
\]

and a probability that \( N(t) \) events occur during time \( t \) of

\[
p_{\rho}(N(t) = m) = \frac{[\Lambda(t)]^m \exp[-\Lambda(t)]}{m!} \tag{2}
\]

Using equations (1) and (2), the mean number of events is found to be

\[
\overline{N(t)} = E_{\lambda}[\rho[N(t)]] = E_{\lambda}[\Lambda(t)] = \int_0^t E_{\lambda}([\Lambda(T)]) \, dT = \rho t \tag{3}
\]

where \( E_{\lambda}(\cdot) \) denotes the operator that takes the expectation value of its argument over the set of rates \( \lambda \), \( E_{\lambda}(\cdot) \) denotes the operator that takes the expectation value over the point process defined by equation (2), and the overbar denotes the average over both the rates and the point process. Similarly, the variance in the number of events is

\[
\text{Var}[N(t)] = \left[ \overline{N(t)} - \overline{N(t)} \right]^2
\]

\[
= \rho t + 2 \int_0^t \langle t - t' \rangle \Gamma(t') \, dt' \tag{4}
\]

and the covariance between the numbers of events in two adjacent time intervals is (Takahata 1991)

\[
\text{Cov}[N(t), M(t)] = \left[ \overline{N(t)} - \overline{N(t)} \right] \left[ \overline{M(t)} - \overline{M(t)} \right] = \int_0^t \langle t - t' \rangle \Gamma(t') \, dt' \tag{5}
\]

where \( M(t) \) is the number of events in the interval of duration \( t \) that is adjacent to the interval of \( N(t) \). The index of dispersion is defined as the ratio of the variance (eq. 4) to the mean (eq. 3):

\[
I(t) = \frac{\text{Var}[N(t)]}{\overline{N(t)}} = 1 + \frac{2}{\rho t} \int_0^t (t - t') \Gamma(t') \, dt'. \tag{6}
\]

The constant rate of a homogeneous Poisson process is uncorrelated in time (\( \Gamma(\tau) = 0 \) for \( \tau > 0 \)), so equations (3–5) imply that the variance equals the mean and that the covariance is zero. From equation (6) it is evident that the index of dispersion equals unity for a homogeneous Poisson process. A doubly stochastic Poisson process (Cox and Isham 1980), on the other hand, not only stochastically follows equation (2), but also has a stochastic rate \( \lambda(T) \) that is correlated in time according to \( \Gamma(\tau) > 0 \). Grandell (1976) provides a detailed discussion of doubly stochastic Poisson processes that includes a derivation of their first and second moments.

Since the index of dispersion (eq. 6) is elevated above unity to an extent determined by \( \Gamma(\tau) \), Gillespie (1991) suggested viewing molecular substitutions as the events of a doubly stochastic Poisson process in order to model the overdispersion he found in mammalian sequence comparisons. Then, \( N(t) \) and \( M(t) \) represent the numbers of substitutions that occur in two related species since their time of divergence \( t \) from a common ancestor. Gillespie selected NFDSPPs that caused the index of dispersion (eq. 6) to be independent of the duration of the observational window \( t \) and thus independent of the mean number of substitutions (eq. 3). However, selecting slowly decaying autocovariance functions leads to very different statistics, as is seen in the next section.

**Fractal Doubly Stochastic Poisson Process**

The FDSPP considered presently has long-range correlations in its rates in the sense that the integral of the autocovariance function \( \Gamma(\tau) \) over all times diverges. Such a process can be constructed by postulating an inverse power-law autocovariance function (Abry and Flandrin 1996) such as

\[
\Gamma(\tau) = \text{Cov}_{\lambda}[\lambda(T), \lambda(T + \tau)] = E_{\lambda}([\lambda(T) - E_{\lambda}[\lambda(T)]][\lambda(T + \tau) - E_{\lambda}[\lambda(T + \tau)])
\]

\[
= \frac{C}{(\tau + \tau_0)\beta}, \tag{7}
\]

where \( \tau_0 \) is a positive real constant representing the minimum lag time \( \tau \) before equation (7) has inverse power-law behavior in \( \tau \), where \( \beta \) is an exponent satisfying

\[
0 < \beta < 1, \tag{8}
\]

and where \( C \) is a normalization constant related to \( \tau_0 \) by

\[
C = \frac{\tau_0^\beta \Gamma(0)}{\tau_0^\beta \text{Var}_{\lambda}(\lambda(t))}. \tag{9}
\]

For large values of \( t \) \( (t \gg \tau_0) \), substituting equation (7) into equations (4) and (5), respectively, yields the FDSPP variance,

\[
\text{Var}[N(t)] = \rho t + 2 \int_0^t (t - t') \frac{C}{(t')^\beta} \, dt' = \rho t + \rho k t^{2-\beta}
\]
Therefore, and covariance
\[ \text{Cov}[N(t), M(t)] = \int_0^{2I} (t - |t - t'|) \frac{C}{(r')^b} \, dt' \]
\[ = (2^{1-\beta} - 1) \rho \kappa t^{2-\beta}, \]
where \( \kappa \) is the constant
\[ \kappa = \frac{2C}{(1 - \beta)(2 - \beta)^b}. \]
The index of dispersion (eq. 6) is obtained by dividing equation (10) by equation (3):
\[ I[N(t)] = 1 + \kappa t^{2-\beta} = 1 + \kappa t^p. \]
Here, \( D = 1 - \beta \) and consequently lies in the range \( 0 < D < 1. \)
For sufficiently large time windows \( t \), the index of dispersion increases as a power-law,
\[ I(t) \approx \kappa t^p, \]
and thus, for any arbitrary constant \( \theta \), follows the scaling relation
\[ I(\theta t) = \theta^D I(t). \]
Therefore, \( D \) is called the scaling exponent of the process and is a measure of the amount of correlation in a process over long periods of time \( t \). From equation (16), it is evident that the index of dispersion is fractal, which is to say that it has self-similarity upon scaling. Lowen and Teich (1995) thus call processes following equation (13) fractal stochastic point processes and call \( D \) the fractal dimension. The process developed above is but one example of an FDSPP, which in turn is only one type of a fractal point process, since there are many ways to construct processes that scale in their indices of dispersion (Lowen and Teich 1995).

In viewing molecular evolution as an FDSPP, the events of the process are molecular substitutions, so that for a gene with label \( i \), \( N_i(t) \) and \( M_i(t) \) are the numbers of substitutions occurring in the \( i \)th gene of two organisms during the time \( t \) that elapsed since their divergence from a common ancestor. Since each gene evolves at a different rate, for the \( i \)th gene, let \( \lambda_i(t) \) denote the instantaneous rate of substitution, let \( \rho_i \) denote the mean rate of substitution, and let \( \tau_{0i} \) denote the minimum lag time before the autocorrelation function (eq. 7) approximates an inverse power-law. The fractal dimension \( D = 1 - \beta \) is taken to be a gene-independent parameter of the model. Equation (15), which gives the index of dispersion as a function of divergence time, can be written instead as a function of the mean number of substitutions \( \bar{N}_i \) in the \( i \)th gene using equations (3), (9), and (12):
\[ I(\bar{N}_i) \approx \frac{2\tau_{0i}^\beta \text{Var}_a(\lambda_i)}{(1 - \beta)(2 - \beta)\rho_i} \left( \frac{\bar{N}_i}{\rho_i} \right)^D \]
\[ = \frac{2(\rho_i\tau_{0i})^{1-D} \text{Var}_a(\lambda_i)}{D(D + 1)} \left( \frac{\text{Var}_a(\lambda_i)}{\rho_i^2} \right) \bar{N}_i^D, \]
where the mean number of substitutions is assumed to be large. The argument \( t \) has been dropped, since equation (17) treats the divergence time as a constant and the mean number of substitutions as the dependent variable. We define two dimensionless rate-independent parameters \( \chi \) and \( \delta \), respectively, as the proportionality constant between \( \tau_{0i} \) and the reciprocal of the mean rate,
\[ \chi = \rho_i\tau_{0i}, \]
and the coefficient of variation of the rate,
\[ \delta = \sqrt{\frac{\text{Var}_a(\lambda_i)}{\rho_i}}, \]
Thus, the index of dispersion increases as a power-law (geometrically or algebraically) in the mean number of substitutions:
\[ I(\bar{N}_i) \approx \gamma \bar{N}_i^D, \]
with the constant \( \gamma \) depending only on the three gene-independent parameters of the FDSPP model:
\[ \gamma = \frac{2\chi^{1-D} \delta^2}{D(D + 1)}. \]
It can similarly be demonstrated that the coefficient of variation of the number of substitutions decreases as a power-law in the mean:
\[ C(\bar{N}_i) = \sqrt{\frac{\text{Var}_a(\lambda_i)}{\bar{N}_i}} \approx \frac{1}{\sqrt{\bar{N}_i}} + \frac{\gamma}{\bar{N}_i^{1-D}}. \]
The following section proves that equations (20) and (22) are in agreement with DNA sequence comparisons.

**Fit to Mammalian Sequence Data**

Gillespie (1991) estimated the index of dispersion for substitution counts in humans, rodents, and artiodactyls without assuming equal divergence times or the same substitution rate for each lineage. Using Gillespie’s method, let the index \( i \) label a particular gene, where \( i \) ranges from 1 to the number of genes \( n \), let the index \( j \) label the species, where \( j = 1 \) for humans, \( j = 2 \) for rodents, and \( j = 3 \) for artiodactyls, let \( N_{ij} \) denote the estimated number of substitutions in gene \( i \) from the common ancestor to species \( j \), and let \( d_{jk} \) denote the observed number of site differences between the \( i \)th gene of species \( j \) and the \( k \)th gene of species \( k \) after correcting for multiple substitutions. Figure 1 depicts the relationship between \( N_{ij} \) and \( d_{jk} \):
\[ d_{i1} = N_{i1} + N_{i2} \]
\[ d_{i2} = N_{i2} + N_{i3} \]
\[ d_{i3} = N_{i3} + N_{i1}. \]
FIG. 1.—Three-species tree for a gene of label $i$, in which humans, artiodactyls, and rodents are assumed to share a common ancestor without assuming equal divergence times as in a star phylogeny. The estimated numbers of substitutions between species are used to compute the numbers of substitutions since divergence from a common ancestor, as described in the text.

Table 1

<table>
<thead>
<tr>
<th>Weight Type</th>
<th>Substitution Type</th>
<th>Estimated $\gamma$</th>
<th>Estimated $D$</th>
<th>$D$ 95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonweighted</td>
<td>Synonymous</td>
<td>0.41</td>
<td>0.90</td>
<td>(0.60, 1.21)</td>
</tr>
<tr>
<td>Weighted</td>
<td>Synonymous</td>
<td>0.21</td>
<td>0.59</td>
<td>Not available</td>
</tr>
<tr>
<td>Nonweighted</td>
<td>Nonsynonymous</td>
<td>0.36</td>
<td>0.71</td>
<td>(0.34, 1.07)</td>
</tr>
<tr>
<td>Weighted</td>
<td>Nonsynonymous</td>
<td>0.66</td>
<td>0.45</td>
<td>(0.16, 0.74)</td>
</tr>
</tbody>
</table>

NOTE.—The above parameters were found by minimizing (eq. 36) using the data of Ohta’s (1995) table 3. In all cases, the estimated fractal exponent $D$ is between 0 and 1, as is required for the FDSPP. When weights are used, the strength $D$ of long-range correlations is decreased due to the removal of the correlation caused by lineage effects such as unequal branch times and lineage-dependent rates.

Fig. 2.—Index of dispersion versus mean for weighted nonsynonymous substitutions. The sample correlation coefficient is only 0.162, but, as the text argues, this is not an appropriate measure of correlation in this case. The line is the best fit to the $n = 49$ data points representing the 49 genes analyzed by Ohta (1995).
that Gillespie’s NFDSPP models were compatible with her weighted nonsynonymous results, because she believed the correlation between the index of dispersion and the mean number of substitutions to be insignificant. Her hypothesis of no correlation can be examined using the test for correlation between two random variables $x$ and $y$, assuming that they follow the bivariate normal probability distribution

$$f(X, Y) = (2\pi\sigma_x\sigma_y\sqrt{1 - r^2})^{-1} \times \exp\{-[(X-\mu_x)/(\sigma_x)]^2 - 2r(X-\mu_x) \times (Y-\mu_y)/(\sigma_y) + ((Y-\mu_y)/(\sigma_y))^2]/[2(1-r^2)]\},$$

(29)

as described in standard statistics texts such as that of Devore (1990), where $\mu_x$ and $\mu_y$ are the means of $x$ and $y$, $\sigma_x$ and $\sigma_y$ are the standard deviations of $x$ and $y$, and $r$ is the population correlation coefficient, which is estimated by the sample correlation coefficient

$$R = \frac{n \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \sqrt{n \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}},$$

(30)

where the $n$ pairs $(x_1, y_1), \ldots, (x_n, y_n)$ make up the data sample. The null hypothesis of no correlation ($r = 0$) is rejected at confidence level $\alpha$ if and only if $t_{R,n} \geq t_{a/2,n-2}$ or $T_{R,n} \leq -t_{a/2,n-2}$, where $T_{R,n}$ is the test statistic

$$T_{R,n} = R\sqrt{n - 2}/\sqrt{1 - R^2}$$

(31)

and where $t_{a/2,n-2}$ is the critical value for the $t$ distribution with the first argument designating the confidence level for a one-tailed $t$ distribution and the second argument designating the number of degrees of freedom. In applying this test to Ohta’s (1995, table 3) weighted nonsynonymous substitution data, we let $x_i$ be the estimation of the mean number of substitutions in the $i$th gene (eq. 26), $y_i$ be the estimation of the index of dispersion of the number of substitutions in the $i$th gene (eq. 28), and $n$ equal 49, the number of genes in her analysis. See figure 2 for a plot of $y_i$ versus $x_i$ with its best-fit line. (Figure 3 plots the same points on a log-log scale and is discussed below.) This sample’s correlation coefficient (eq. 30) is 0.162, so assuming that these data are distributed according to the bivariate normal probability distribution (eq. 29), the test statistic (eq. 31) is 1.13, which implies that the hypothesis of no correlation cannot be rejected at the 10% significance level, a result consistent with the NFDSPP. However, two necessary conditions for the data’s following the bivariate normal probability distribution (eq. 29) are that the $x_i$ values are normally distributed and that the $y_i$ values are normally distributed. Figures 4 and 5 show that neither the means nor the indices of dispersion follow a normal (Gaussian) distribution, so the above test cannot be used directly.

Furthermore, a small correlation coefficient between the index of dispersion and the mean does not imply a lack of a relationship, but only a lack of a linear...
relationship. Consequently, we test whether there is a geometric or power-law relationship as suggested by the FDSPP model (eq. 20). Taking the logarithm of both sides of equation (20) yields an equation that is linear in the logarithms of the index of dispersion and mean number of substitutions:

$$\log(I(N)) \approx D \log(N) + \log(\gamma).$$  \hfill (32)

It follows that a linear relationship in the logarithms (eq. 32) necessitates a power-law relationship between the index of dispersion and the mean (eq. 20). Since the logarithms of the means are normally distributed and the logarithms of the dispersion indices are approximately normally distributed, as indicated in figures 6 and 7, if we redefine $x_i$ and $y_i$ as the logarithms of the estimated mean and index of dispersion of weighted nonsynonymous substitutions in the $i$th gene, as plotted in figure 3, then the assumption of a bivariate normal distribution (eq. 29) is plausible. In this case, the sample's correlation coefficient (eq. 30) is 0.418, so the test statistic (eq. 31) is 3.15, which implies that the hypothesis of no correlation can be rejected at the 0.5% significance level. Thus, there is a strong linear relationship between the logarithm of the index of dispersion and the logarithm of the mean (eq. 32), which implies that there is a strong power-law relationship between the index of dispersion and mean, as required by the FDSPP model (eq. 20). The lognormality in numbers of substitutions could have resulted from a multiplicative process (Bickel and West, unpublished data).
Although the above analysis uses equation (15) as the index of dispersion for the FDSPP, equation (13) could be used instead. The latter relation predicts that the index of dispersion minus unity increases as a power-law in the mean number of substitutions, and thus that

$$\log[I(N_i) - 1] = D \log N_i + \log \gamma,$$

(33)

in analogy with equation (32). Since the logarithm of a negative number is undefined, only those 39 genes with index of dispersion estimates greater than unity can be analyzed using equation (33). In this case, again using the normality test and the t-test, the hypothesis of no power-law correlation between \(I(N_i) - 1\) and \(N_i\) can be rejected at the 0.5% significance level, in agreement with the results of the above analysis that considered all 49 genes. This supports using equation (15) as an approximation of equation (13) in the case of nonsynonymous substitutions.

Even though the analyses presented herein rely on assuming that substitutions have lognormal distributions and that the bias in the standard index-of-dispersion estimator is negligible, the same result can be obtained without these assumptions. Bickel and West (1998) find, using Spearman’s nonparametric correlation coefficient and Bulmer’s (1989) unbiased estimator of the index of dispersion, that for mammals, the correlation of the index of dispersion with the mean number of substitutions is significant for weighted nonsynonymous substitutions at the 1% level and for weighted synonymous substitutions at the 10% level.

The positive power-law correlation found between estimates for the index of dispersion and the mean number of substitutions conflicts with the prediction of Gillespie’s NFDSPP models that the index of dispersion is independent of the mean, but agrees with the prediction of the FDSPP model (eq. 20). Thus, the FDSPP model provides a better fit to the pattern of mammalian non-synonymous substitutions than do previously proposed NFDSPP models.

### Estimation of Model Parameters

In order to use Ohta’s (1995, table 3) data to estimate the scaling exponent \(D\) of the FDSPP, we assume a multiplicative error \(E_i\) in the estimation \(y_i\) of the index of dispersion of the number of substitutions (eq. 28) and introduce \(D'\) and \(\gamma'\) as estimators of \(D\) and \(\gamma\) of equation (20), such that

$$y_i = \gamma'x^D E_i,$$

(34)

the logarithm of which yields the equation for a line of slope \(D'\) and intercept \(\log \gamma'\) plus an error term \(\log E_i:\n
$$\log y_i = D' \log x_i + \log \gamma' + \log E_i.$$  

(35)

Parameter estimates \(D'\) and \(\gamma'\) are found by minimizing the sum of squared errors over \(n\) genes (\(n = 49\) for Ohta’s [1995] data):

$$\text{SSE} = \sum_{i=1}^{n} (\log E_i)^2$$

$$= \sum_{i=1}^{n} [\log y_i - (D' \log x_i + \log \gamma')]^2.$$  

(36)

Table 1 lists values of \(D'\) and \(\gamma'\) for weighted and nonweighted and synonymous and nonsynonymous substitutions as determined by minimizing equation (36) through standard linear regression. The 95% confidence intervals for \(D\) given in table 1 depend on a \(t\) distribution, which assumes the normality of the transformed standardized residuals (transformed for the logarithmic scale from Devore [1990]),

$$e_i = \frac{\log E_i - \log \gamma'}{s \sqrt{1 - \frac{1}{n} - \frac{(x_i - \mu_x)^2}{\sum_{j=1}^{n} (x_j - \mu_x)^2}}},$$  

(37)

where \(\mu_x\) is the mean of all \(n\) of the \(x_i\) values and \(s\) is the estimated transformed standard deviation of the \(y_i\) values, given by

$$s = \sqrt{\frac{\sum_{i=1}^{n} (\log E_i)^2}{n - 2}}.$$  

(38)

Note that if the transformed standardized residuals \(e_i\) are normally distributed, then the errors in the indices of dispersion \(y_i\) are lognormally distributed. West and Bickel (1998) use the test described in the appendix to determine whether the transformed standardized residuals (eq. 37) are normally distributed. In all cases except that of weighted synonymous substitutions, the result is that assuming a normal distribution of transformed standardized residuals (lognormal error) is plausible, and thus the confidence intervals of table 1 have meaning.

### Alternate Models of Molecular Evolution

We have demonstrated that the FDSPP model of molecular evolution is compatible with mammalian sequence comparisons, unlike certain previously proposed variations of the Poisson process model. However, other point process models also agree with the high index of dispersion that increases with the mean number of substitutions. Examples of alternate models include other fractal processes as well as other orderly point processes.

Lowen and Teich (1995) describe a number of different types of fractal point processes, all of which follow equation (13) and thus predict the observed power-law correlation between the index of dispersion and the mean number of substitutions according to equation (20). Such point processes include the FDSPP and the fractal renewal point process, defined as per Lowen and Teich (1993) and applied to molecular evolution by Bickel and West (1998). In addition to fractal point processes, fractal diffusion processes also make good models of molecular evolution (West and Bickel 1998). Unlike most nonfractal models, such as the NFDSPPs, fractal models have slowly decaying rate autocorrelations, as would be expected under the assumption that natural selection causes most molecular substitutions (see Conclusions).

Several models besides fractal models also agree with an elevated index of dispersion that increases with
the mean number of substitutions. Gillespie (1991, p. 130) pointed out that all processes in which no more than one event occurs at a time (orderly or regular processes) have an index of dispersion close to unity for sufficiently short counting times. This is because only 0 or 1 event can occur during a sufficiently small time window, so that the number of events becomes a Bernoulli sequence, the index of dispersion of which is

$$I_{\text{Bernoulli}}(p) = \frac{p(1 - p)}{p} = 1 - p.$$  

(39)

where $p$ is the probability that an event will occur during a time interval $t$ (Turchet and Teich 1996). As $t$ approaches 0, $p$ also approaches 0, which implies that the index of dispersion approaches 1. The index of dispersion decreases as $t$ and $p$ increase, until the Bernoulli sequence becomes a poor approximation of the point process. After that, the index of dispersion approaches an asymptotic value for all processes without long-range correlations. This is because the rate autocovariance function $\Gamma(\tau)$ of such processes goes to zero for sufficiently large lags $\tau$. According to equation (6), this has the result that, in NFDSPPs, the index of dispersion does not depend on the counting time $t$ for sufficiently large $t$. Therefore, orderly processes without long-range correlations but with an equilibrium index of dispersion greater than unity predict an increase in the index of dispersion with time for intermediate times between Bernoulli behavior ($I \leq 1$) and equilibrium ($I > 1$). This agrees with the significant correlation between the estimates of the index of dispersion and mean numbers of nonsynonymous substitutions that was demonstrated previously. However, this correlation is only significant for relatively short times in nonfractal models, whereas for fractal point processes, equation (13) holds for times spanning orders of magnitude. Models with long-range correlations, such as the FDSPP and certain other fractal models, have increasing indices of dispersion for all sufficiently large counting times. Gillespie’s limiting case of an NFDSPP, on the other hand, predicts no correlation between the index of dispersion and the mean number of substitutions. Another process with long-range correlations is that in which each lineage evolves at a constant rate that is assigned randomly at the time of divergence from the common ancestor. In this case, the index of dispersion increases linearly in time (Takahata 1987), which implies that the substitution coefficient of variation is constant. This is incompatible with Ohta’s (1995) observation that the coefficient of variation decreases with an increasing mean number of substitutions, as fractal processes predict (eq. 20). Therefore, fractal models have the desirable property that for a wide range of counting times, the index of dispersion increases sublinearly, a characteristic not shared by most nonfractal models.

The FDSPP is the simplest Poisson process model that fits the mammalian sequence data and that has fractal characteristics, and is therefore a suitable replacement for the homogeneous Poisson process for applications such as testing for the presence of lineage-de-
used as a basis for deriving estimates of the number of substitutions between two sequences.

The Markovian master equation (40) can be modified for the FDSPP by allowing the rates to depend on time:

$$\frac{dP_i(t)}{dt} = \sum_{k=\text{A, G, C, T}} u_{ik}(t)P_k(t),$$

$$i, j = \text{A, G, C, T}. \tag{43}$$

This equation can be used in place of equation (40) to derive corrections for multiple and back substitutions. Owing to the similarity of equations (40) and (43), the resulting estimates of numbers of substitutions are the same whether or not the rates change in time. This is true regardless of the assumptions made about differences in rates for transitions and transversions and about rate variation across sites.

To illustrate the fact that standard estimates of numbers of substitutions do not rely on rate constancy, we will derive the correction for multiple and back substitutions assuming variable rates and the same rate of substitutions between two sequences.

The probability distribution of the number of substitutions between two lineages separated by twice the time $t$ after divergence time $t$ equals the probability that the sites would differ if each site were chosen randomly (uniform on A, G, C, T) and if a site changed to a random nucleotide at least once, with equal probabilities of all nucleotides. The probability that two random nucleotides differ by chance is $3/4$, and the probability of at least one event equals unity minus the probability of no events, so that

$$P(2t) = 3\left\{1 - p_k \left[\frac{N'(2t)}{S} = 0\right]\right\}, \tag{51}$$

where $N'(2t)$ is the number of nucleotides changing randomly in time $2t$. The rate of a site changing to any of four random nucleotides is $4/3$ the rate of a site changing to one of the three nucleotides other than its current nucleotide; thus, the probability distribution of the number of changes $N'(2t)$ to a uniformly distributed random nucleotide per site is (using eq. 49),

$$p_k \left[\frac{N'(2t)}{S} = m\right] = \frac{\left[\frac{4}{3}K(2t)\right]^m \exp\left[-\frac{4}{3}K(2t)\right]}{m!}. \tag{52}$$

Substituting equation (52), with $m = 0$, into equation (51) yields

$$P(2t) = 3\left\{1 - \exp\left[-\frac{4}{3}K(2t)\right]\right\}. \tag{53}$$

Equations (50) and (53) may be combined to obtain

$$E_p \left[\frac{N(2t)}{S}\right] = K(2t) = 2K(t). \tag{50}$$

The probability $P(2t)$ that two sequences differ at a site after divergence time $t$ equals the probability that the sites would differ if each were chosen randomly (uniform on A, G, C, T) and if a site changed to a random nucleotide at least once, with equal probabilities of all nucleotides. The probability that two random nucleotides differ by chance is $3/4$, and the probability of at least one event equals unity minus the probability of no events, so that

$$P(2t) = 3\left\{1 - p_k \left[\frac{N'(2t)}{S} = 0\right]\right\}, \tag{51}$$

where $N'(2t)$ is the number of nucleotides changing randomly in time $2t$. The rate of a site changing to any of four random nucleotides is $4/3$ the rate of a site changing to one of the three nucleotides other than its current nucleotide; thus, the probability distribution of the number of changes $N'(2t)$ to a uniformly distributed random nucleotide per site is (using eq. 49),

$$p_k \left[\frac{N'(2t)}{S} = m\right] = \frac{\left[\frac{4}{3}K(2t)\right]^m \exp\left[-\frac{4}{3}K(2t)\right]}{m!}. \tag{52}$$

Substituting equation (52), with $m = 0$, into equation (51) yields

$$P(2t) = 3\left\{1 - \exp\left[-\frac{4}{3}K(2t)\right]\right\}. \tag{53}$$

Equations (50) and (53) may be combined to obtain

$$E_p \left[\frac{N(2t)}{S}\right] = -\frac{3}{4} \log \left[1 - \frac{4}{3}P(2t)\right], \tag{54}$$

which can be used to estimate $N(2t)$ by estimating $P(2t)$ with the proportion of sites that differ between genes of different species. This, of course, is the correction for multiple and back substitutions for the one-parameter model of Jukes and Cantor (1969).

**Conclusions**

We have demonstrated that the FDSPP model agrees with Ohta’s (1995) mammalian nonsynonymous substitution data much better than do Gillespie’s (1991) NFDSPP models. The FDSPP model has the episodic property, discussed by Gillespie (1991, pp. 132–139), that protein evolution occurs in bursts or clusters of substitutions separated by periods of relative inactivity. While Gillespie cites time-varying substitution rates as evidence for positive natural selection, Ohta (1995) argues that her nearly-neutral theory also agrees with episodic evolution. The relationship between the FDSPP model and the various selection and neutral theories that have been proposed to date has yet to be investigated.

Although the FDSPP model described herein has a mean rate $\rho$, that rate does not characterize the process
as does the mean rate of a Poisson process. Instead, the
FDSSP model indicates that molecular evolution has no
fundamental timescale but, rather, that it has all time-
scales. In other words, substitutions do not all occur at
roughly some characteristic rate, but can occur very
quickly, very slowly, or at any intermediate rate; the
instantaneous rate \( \lambda(T) \) is often very much higher or
very much lower than the average rate \( \rho \). Rather than
being completely specified by its mean rate, an FDSSP
is also characterized by its fractal dimension \( D = 1 - \beta \),
since that parameter determines how strongly the pro-
cess is correlated in time (eq. 7). High values of \( D \) imply
that the instantaneous rate \( \lambda(T) \) of substitutions at time
\( T \) depends strongly on the instantaneous rates at all pre-
vious times, including rates from millions of years ago,
while lower values of \( D \) weaken those long-range cor-
relations. Like the index of dispersion (eq. 16), the rate
autocovariance function (eq. 7) for \( \tau \gg \tau_0 \) has self-sim-
ilarity upon scaling:

\[
\Gamma(\theta \tau) = \theta^{-\beta} \Gamma(\tau).
\]

(55)

Thus, to paraphrase Bassingthwaighte, Liebovitch, and
West (1994), the correlation between neighbors of 100
Myr is the same as the correlation between neighbors of
10 Myr (West and Bickel 1998).

The autocorrelation of the rate of molecular evolu-

tion over a large range of timescales could result from
selection pressures that persist for widely varying du-
rations, assuming that natural selection causes most non-
synonymous nucleotide substitutions in DNA. We pres-
ent examples of natural selection in mammals acting
during timescales roughly on the order of \( 10^4 \), \( 10^5 \), and
10\(^9\) years under Darwinian theory. First, the prevalence
of the sickle-cell type of beta hemoglobin in regions of
high frequencies of malignant malaria caused by the
mosquito *Plasmodium falciparum* evidences natural se-
lection in human evolution over thousands of years. Se-
lection pressures occurring during a much larger time-
scale caused the evolution of modern man from more
apelike men, according to the multiregional theory of
human origins. In a period of a few hundred thousand
years, the brain size of *Homo erectus* gradually in-
creased due to the selective advantages of a larger brain
until *Homo erectus* populations evolved into *Homo sap-
ien* populations according to this theory. On a still larger
timescale, the global environmental changes that result-
ed from the strike of the Alvarez asteroid in Yucatán
about a billion years ago persist even to the present. This
cataclysm apparently caused the extinction of the di-
inosaur and provided a number of new niches to which
mammals adapted by means of natural selection. The
evolution in today’s mammals is affected by environ-
mental selective pressures that were initiated by the me-
eteor impact, so the rate of molecular evolution in mod-
ern mammals is highly correlated with the rate in mam-
mals from a billion years ago. Therefore, there are both
short-range and long-range correlations in the rate of
mammalian molecular evolution. Long-range correla-
tions imply an inverse power-law autocorrelation func-
tion, so a fractal model such as FDSSP is an excellent
choice of a model that is in agreement with Darwinian

selection pressures that act on all timescales. According
to this interpretation, the stochastic nature of the model
could correspond to the randomness of the mutations on
which natural selection acts.

The agreement of the FDSSP model of molecular
evolution with standard Darwinian theory sets it apart
from some other fractal concepts of evolution. For ex-
ample, the genomic potential hypothesis of Schwabe
(1990a, 1990b, 1994) suggests that adaptation emerges
by deterministic processes rather than through chance
and selection. Similarly, the FDSSP rate \( \lambda(t) \), viewed as
a chaotic process, would imply deterministic evolution.
In this case, the stochastic nature of the numbers of sub-
stitutions results in part from a sensitive dependence on
the initial conditions of the evolving population. There-
fore, the FDSSP can be interpreted in terms of chaotic
determinism or Darwinian chance and natural selection.

Further research is needed to determine how re-
placing the homogeneous Poisson process with the
FDSSP will impact the evaluation of the lineage-depen-
dence of rates, the examination of the relative roles of
selection and drift in evolution, the computation of con-
fidence intervals for estimated divergence times, and
other model-dependent investigations in molecular bi-
ology.

**Acknowledgment**

We would like to thank the Office of Naval Re-
search for the partial support of this work.

**APPENDIX**

**Test of Whether a Sample is Normally Distributed**

The following test, from Devore (1990), determines
whether a sample of \( n \) values \( U_i \) follow a normal dis-
tribution, where \( i \) runs from 1 to \( n \). Relabel the values
\( U_i \) such that they are in order from the smallest value,
\( U_{i_1} \), to the largest value, \( U_{i_n} \). The \( z \) percentiles, \( z_c \), for a
normal distribution are

\[
z_c = \Phi^{-1} \left( i - \frac{3}{8} \right) \left( \frac{n + 1/4}{n + 1/4} \right).
\]

(56)

where \( \Phi \) is the cumulative standard normal distribution
function, defined by

\[
\Phi(Z) = (2\pi)^{-1/2} \int_{-\infty}^{Z} e^{-z^2} \, dz.
\]

(57)

A normal probability plot consists of a graph of the
sample values \( U_i \) versus the normally distributed values
\( z_i \). Examples are displayed in figures 4–7. Let \( r \) be the
sample correlation coefficient (eq. 30) for the \( n \) pairs
(\( z_i, U_i \)). The hypothesis that the \( U_i \) values follow the normal
distribution is rejected when \( r \leq c_\alpha \), where \( c_\alpha \) is the
critical value at significance level \( \alpha \). Devore (1990)
provides a list of critical values for various values of \( n \) at
the 1%, 5%, and 10% significance levels.

**LITERATURE CITED**

Abram, P., and P. Flandrín. 1996. Point processes, long-range
dependence, and wavelets. Pp. 413–437 in A. Aldroubi

Marcy K. Uyenoyma, reviewing editor

Accepted April 14, 1998