Testing the Covarion Hypothesis of Molecular Evolution

Michael M. Miyamoto* and Walter M. Fitch†

*Department of Zoology, University of Florida; and †Department of Ecology and Evolutionary Biology, University of California

The covarion hypothesis of molecular evolution states that the fixation of mutations may alter the probability that any given position will fix the next change. Tests of this hypothesis using the divergence of real sequences are compromised because models of rate variation among sites (e.g., the gamma version of the one-parameter equation) predict sequence divergence values similar to those for the covarion process. This study therefore focuses on the extent to which the varied and unvaried codons of two well-diverged taxa are the same, because fewer are expected by the covarion hypothesis than by the gamma model. The data for these tests are the protein sequences of Cu, Zn superoxide dismutase (SOD) for mammals and plants. Simulation analyses show that the covarion hypothesis makes better predictions about the frequencies of varied and unhit positions in common between these two taxa than does the gamma version of the one-parameter model. Furthermore, the analysis of SOD tertiary structure demonstrates that mammal and plant variabilities are distributed differently on the protein. These results support the conclusions that the variable and invariable codons of mammal and plant SODs are different and that the covarion model explains the evolution of this protein better than the gamma version of the one-parameter process. Unlike other models, the covarion hypothesis accounts for rate fluctuations among positions over time, which is an important parameter of molecular evolution.

Introduction

The covarion (concomitantly variable codon) hypothesis states that at any particular point in the evolutionary history of a protein, only a fraction of its amino acid positions are free to vary (Fitch and Markowitz 1970; Fitch and Ayala 1994a, 1994b). Codons of a protein are separable into variable and invariable classes, with the latter unable to accept amino acid replacements due to functional and selective constraints. The possibility of invariable sites is not unique to this hypothesis, since their existence was considered by the earliest models of protein evolution (Margoliash and Smith 1965; Uzzell and Corbin 1971). However, the covarion hypothesis remains unique in its proposal that the memberships of the variable and invariable pools change over time (Fitch 1971a). As replacements occur among covarions, members of this class are exchanged for positions that were previously invariable. These exchanges explain how protein sequences can differ considerably among distant taxa even though only a limited number of positions are free to vary within a particular lineage.

In more specific terms, the covarion hypothesis proposes the existence of three different categories for the n amino acid positions of a protein: the covarion pool with its set of currently variable sites (c); the class for presently invariable codons that can potentially become variable (ti, for temporarily invariable); and the category for permanently invariable positions (pi). Collectively, the c and ti classes comprise the total pool of potentially variable positions (pv). Conversely, at any moment in time, ti plus pi codons are invariable and are unavailable for change. The rate at which codons in the covarion class are exchanged with those in the ti pool is referred to as the persistence of variability (v) (Fitch 1971a). The covarion hypothesis treats these exchanges as if made on a one-to-one basis (i.e., one member of the covarion pool for one representative in the ti class), although such rigid numerical equality is not biologically realistic. The hypothesis also assumes that the rate of codon interchange is related to the frequency of replacements in the protein (see below).

The functional and selective constraints that characterize the three classes and regulate the rate of exchange between site variability and invariability can best be un-
derstood by reference to specific examples. The initiating methionine of proteins provides the best example of an amino acid position that is most likely permanently invariable, because of its essential role in polypeptide synthesis. The reasonableness of interchange between site variability and invariability is illustrated by bovine ribonuclease A (Fitch and Markowitz 1970; Fitch and Ayala 1994a, 1994b), in which the positively charged arginine at position 39 is considered invariable, because of its functional importance as a shield for the positively charged lysine at site 41 in the active site from the negative aspartate at position 38. Replacement of aspartate 38 with a nonnegative amino acid (as in the rat) relieves arginine 39 of its shielding responsibility, thereby allowing it now to vary. Interestingly, no combination of amino acids at varied positions 38 and 39 has yet been found among mammals to contradict the possibility that arginine 39 is invariable as long as site 38 is occupied by a negatively charged amino acid.

The most recent investigations of the covarion hypothesis and its applicability to actual protein sequences have focused on the molecular evolution of Cu, Zn superoxide dismutase (SOD, EC 1.15.1.1) (Fitch and Ayala 1994a, 1994b). In these studies, 67 SOD sequences were compared among bacteria and higher organisms, which resulted in the following estimates of parameters for metazoan evolution under this model: 162 total codons, with 118 potentially variable and 44 permanently invariable among the metazoans, including those only present in bacteria; a covarion pool of 28 codons; a persistence of variability of 0.01 (i.e., one exchange of a currently variable codon in the covarion class with a temporarily invariable position for every 100 replacements); an average of 2.5 amino acid alternatives per variable site ($a$); and 0.3 replacements per million years ($\lambda$). Simulations with these parameters led to predicted estimates of SOD differences that agreed well with observed values (fig. 1). This agreement supported the conclusion that the covarion hypothesis is a reasonable explanation of SOD evolution.

Those investigations did not consider any alternative models of protein change, as their stated purpose was to test the reasonableness of the covarion hypothesis. An earlier study of a subset of the 67 SODs (Kwiatowski et al. 1991) documented a large decrease in its apparent replacement rate over time that was not easily accounted for by standard models of protein evolution (Lee et al. 1985; Ayala 1986). As a result, rate variation among positions was hypothesized to explain the puzzling behavior of the protein. The occurrence of both rapidly and slowly evolving codons is of primary importance to SOD evolution according to this alternative explanation.

Substitution and replacement rates are related to the gamma distribution in the most extensively developed and used models of site-to-site variation in $\lambda$ (Nei and Gojobori 1986; Jin and Nei 1990; Nei 1991; Tamura and Nei 1993; Yang 1993; Rzhetsky and Nei 1994; Tateno et al. 1994; Huelsenbeck 1995; see also Olsen 1987; Gouy and Li 1989). Protein differences are expected to increase over time ($t$) under the one-parameter model (Jukes and Cantor 1969) with rate variation among sites according to the equation

$$p = b\{1 - [(a/(a+2\lambda/b)]^2\},$$

(1)

where $p$ is the proportion of different codons between two sequences, $a$ is the shape and scale of the gamma distribution for $\lambda$, and $b$ is ($1 - \text{the probability of chance similarity between the sequences}$) or 19/20 for amino acid data (Nei and Gojobori 1986). That rate heterogeneity among sites is a viable alternative to the covarion hypothesis of SOD evolution is shown in figure 1. The optimized fit of equation (1) to the real data results in predicted values with a smaller sum-of-squared differences for the gamma process ($SSD = 0.009$) than for the covarion hypothesis ($0.015$). As the parameters of
Fig. 2.—Phylogenetic tree for the 14 mammal and plant sequences of Cu, Zn SOD used in this study (Fitch and Ayala 1994a, fig. 1). Branch lengths (replacement substitutions from back translations of the proteins) are corrected for parallel and back changes by gamma eq. (2) with $a = 0.27$ and 0.55 (in parentheses). Branches are drawn proportional to the set of corrected lengths with $a = 0.27$. Common names and GenBank accession numbers for the SOD sequences are given for each species in brackets.

The covarion model were not optimized to the observed differences (Fitch and Ayala 1994a), care must be taken not to overinterpret these results. However, the evidence in figure 1 clearly shows that the gamma version of the one-parameter model is at least a viable alternative to the covarion hypothesis of SOD evolution, given available data.

Additional tests are needed to distinguish between these processes. One possible approach concentrates on the different predictions of the two about the extent to which the variable and invariable codons of one group are the same as those in another (Fitch 1971b). Homologous positions in different lineages share the same rate in the gamma model, despite varying among sites. The identity of rapidly and slowly evolving codons is therefore fixed among groups. In contrast, the divergent nature of the covarion process leads to different pools of variable and invariable codons among lineages. Thus, a position free to vary in one clade may be currently invariable in another. Given similar amounts of change, the gamma model expects greater agreement between the varied and unhit codons of different taxonomic groups than does the covarion hypothesis.

The significance of the covarion process to SOD evolution is tested with this approach. The predictive power of this hypothesis is compared to that of the gamma model by simulations concerned with the identities of variable and invariable codons in mammals and plants. These simulations are followed by an analysis of group variability on the tertiary structure of the protein. Taken together, these tests reveal that the variable and invariable positions of mammals and plants are different, and the main features of SOD evolution can be explained by the covarion hypothesis of molecular evolution.

**Methods**

Seven mammal and seven plant sequences of Cu, Zn SOD were selected from the available 67 (fig. 2). These 14 were chosen because: they belonged to two major lineages (mammals and plants); their individual subtrees (as determined from parsimony analysis; see below) were identical in shape and were similar in total length (101.7 versus 109.7 uncorrected changes, respectively); and they were represented by relatively large numbers of varied and unhit sites, both within and between groups. The alignment and phylogeny of Fitch and Ayala (1994a, fig. 1) was accepted for these 14 sequences. Their phylogeny was determined from a parsimony analysis of all 67 SODs, but with certain groups in the input topology for branch swapping fixed from
the start as "correct." Determinations of "correct" were based on the strength of support from independent (i.e., non-SOD) data for the individual groups. These evaluations were undertaken by Fitch and Ayala (1994a) to enhance the chances of obtaining an accurate estimate of the true sequence phylogeny.

The previous fit of gamma equation (1) to observed SOD differences was based on the questionable assumption that all 20 amino acids are permissible at every codon position (fig. 1). Although usually accepted for protein data, this assumption fails to recognize that many variable codons are limited to fewer than 20 alternatives, because of structural and functional constraints for acidic, basic, polar, nonpolar, and aromatic residues (Dayhoff 1978). The average number of amino acid alternatives is therefore more likely to be similar to that for nucleic acids than to its maximum value (Fitch and Ayala 1994a). That $a = 4$ is at least as reasonable an estimate as $a = 20$ for mammal and plant SODs is supported by their best fits of equation (1) to observed differences (SSD = 0.009 for both values of $a$, with $a = 0.27$ and $\lambda = 0.00158$ for the former; fig. 1). Thus, $b = 3/4$ was accepted for the gamma and one-parameter processes (see below) rather than $b = 19/20$, which is normally assumed by these models for protein sequences (Jukes and Cantor 1969; Rzhetsky and Nei 1994). Protein evolution under these conditions becomes equivalent to that for nucleic acids with their four alternative bases per nucleotide position.

Minimum estimates of branch lengths for the accepted phylogeny of the 14 SODs were determined by back translation of the amino acid sequences into mRNA codons according to the genetic code and subsequent optimization of the inferred mRNA codons onto the topology itself (Fitch 1971c; Fitch and Farris 1974). These counts of replacement substitutions were corrected for parallel and back changes by multiplying each by the gamma equation (Rzhetsky and Nei 1994):

$$d = ab[(1-p/b)^{-1/a} - 1],$$

(2)

where $p$ is the proportion of uncorrected substitutions to total sequence length for a descendant and its immediate ancestor, $d$ is the corrected number of replacement substitutions per position, $b$ is 3/4 for protein sequences with $a = 4$, and $a = 0.27$, which is its optimal value under these conditions according to the best fit of formula (1) to observed SOD differences (fig. 1). Equation (2) was chosen to correct the branch lengths to increase the chances of success by the gamma model in the final simulations (see below).

Three different models of protein sequence evolution were examined. The first was the standard one-parameter model with its uniform rate of amino acid replacements for all codons (Jukes and Cantor 1969). The other two were the covarion hypothesis (Fitch and Ayala 1994a, 1994b) and the gamma version of the one-parameter process (Nei and Gojobori 1986; Nei 1991). These three processes were hereafter referred to as the one-parameter, covarion, and gamma models, respectively.

These models were compared in three successive series of simulations designated I, II, and III. The goals of simulations I and II were to offer insights into the general behavior of each process. The simulated predictions of the three models were then tested against the actual results for mammal and plant SODs in simulations III.

Simulations I were of two protein sequences with $a = 4$, diverging under the covarion and one-parameter models, with the following conditions: total sequence length fixed at the relatively large value of 2,000 to minimize stochastic error; evolutionary rate set to one replacement substitution per unit time; equal frequencies for the four alternatives; identical probabilities for the 12 replacement substitutions; and observed differences (expressed as proportions of $n$) calculated for the two proteins every 40 substitutions for up to 5,000 changes. Covarion size in these runs was placed at 300 or 1,000 amino acids (0.15 or 0.50 of $n$, respectively), whereas $v$ was set to 0.0, 0.5, 0.9, or 1.0.

The covarion, gamma, and one-parameter models were all included in simulations II, with the frequencies of the four alternatives and probabilities of the 12 replacement substitutions as in I. Total sequence length was set to 1,520 or 10 times the number of aligned positions for mammal and plant SODs (see below). Fourteen protein sequences of this length were generated according to the phylogeny of mammals and plants (fig. 2). The gamma-corrected branch lengths with $a = 0.27$ were used as well, except that each was multiplied by 10 as was $n$. Covarion size was fixed in these simulations to 0.18 (280 positions or 10 times the estimate of $c$ by Fitch and Ayala (1994a) or 0.50 (760 sites), whereas $v$ was varied from 0.0 to 1.0 in increments of 0.1. The frequency of permanently invariable sites was set in these runs to 0.22 of $n$ or 340 (1,520 minus 10 times their estimate of $pv$). Values of $a$ were fixed in the gamma simulations at 0.50, 0.75, 1.00, 1.50, 2.00, and 5.00, in addition to 0.27, which is its optimal value according to the best fit of equation (1) to observed SOD differences (fig. 1).

The identities of varied and unvaried codons in one group were compared in simulations II to those of the other clade. The seven sequences corresponding to
mammals comprised group M, whereas the other seven belonged to clade P for plants. These group designations were capitalized when referring to their varied sites. Following these conventions, four counts were made: FMP, frequency of sites varied in both “mammals” and “plants”; FMp, frequency of positions varied in “mammals” but not “plants”; FmP, frequency of positions changed in “plants” but not “mammals”; and Fmp, frequency of sites unhit in both. Two summary statistics were calculated from these frequencies: the fraction of sites changed in either or both groups (%Varied = \[ \frac{FMP + FMp + FmP}{\text{total sequence length}} \]); and the proportion of FMP to varied positions (%FMP = FMP/\[ \frac{FMP + FMp + FmP}{\text{total sequence length}} \]).

Calculations of FMP, FMp, FmP, and Fmp, along with their summary statistics, were also made in simulations III, except that the variables for the covarion, gamma, and one-parameter models were chosen with respect to the mammal and plant SODs themselves. Protein sequences of 152 codons were used in III, since the alignment of mammal and plant SODs was of this length after the removal of positions missing in either or both groups. The phylogeny of figure 2 was used along with its gamma-corrected branch lengths at \( a = 0.27 \). Other parameters in the covarion simulations followed Fitch and Ayala (1994a) including their estimate of \( a = 2.5 \). Additional variables in the gamma (with \( a = 0.27 \)) and one-parameter models were set as in simulations II. Thus, \( a \) was left equal to four in these runs. Comparisons of simulated results to the actual mammal and plant frequencies were accomplished in each case by the \( \chi^2 \) test with three degrees of freedom.

Other values of \( a \) were also systematically tried in the gamma simulations to improve the fit of this model to observed mammal and plant frequencies. The minimum estimates of branch lengths were recorrected every time with the current value of \( a \) and equation (2) to ensure that these parameters were always consistent. A second round of simulations was then conducted for the covarion and one-parameter models with the final set of gamma-corrected branch lengths for the optimal value of \( a \). All other parameters in these runs were fixed as in the previous round of simulations III with \( a = 0.27 \). These additional trials were performed for the three approaches to maximize further the chances of success by the gamma model.

Simulations under the covarion model were accomplished with the FORTRAN program SIMTREES (Fitch and Yc 1991). The gamma and one-parameter simulations were completed with the C program GENSEQ, written by N. Takezaki in the laboratory of M. Nei.

The tertiary structure of bovine Cu, Zn SOD has been determined at 2-Å resolution (Tainer et al. 1983). Its arrangement has been conserved in humans (Getzoff et al. 1989) as presumably in fruitflies (Kwiatowski et al. 1992) and bacteria (Bannister and Parker 1985). This model was followed in the comparisons of mammal and plant variabilities to SOD structure.

Results and Discussion
Simulations I and II

The results of simulations I show that sequence differences accumulate over time more slowly under the
covarion hypothesis than the one-parameter model (fig. 3). This decline in the apparent rate becomes more pronounced as covarion size decreases and/or persistence increases. Sequence evolution under the one-parameter model is defined by the well-known equation (Jukes and Cantor 1969)

\[ p = b(1 - e^{-2x/b}) \]  

where \( b = 3/4 \) for proteins with \( \alpha = 4 \). Given enough time and with \( b = 3/4 \), sequences evolving under equation (3) will reach an asymptote of 0.75. In addition to the one-parameter curves, this equilibrium will also be reached eventually by the lines for \( v = 0.0, 0.5, \) and 0.9, although more slowly (fig. 3). In contrast, the curves for \( v = 1.0 \) approach their asymptotes early in the simulations. Sequence evolution when \( v = 1.0 \) is defined by the following modified version of (3):

\[ p = bx(1 - e^{-2x/b}) \]  

where \( x \) is the proportion of sites free to vary (Palumbi 1989). Variable sites in the covarion pool are not exchanged for other positions under these conditions, thereby effectively reducing overall sequence length to \( c \). The asymptotes for the two curves with \( v = 1.0 \) are therefore 0.125 and 0.375, which are the products of their covarion pools (0.15 and 0.50 of \( n \)) times \( b = 3/4 \), respectively.

Apparent rate slowdowns, as represented by the intermediate curves in figure 3 (\( v = 0.0, 0.5, \) and 0.9), are typical of what other authors have found puzzling about the evolution of SOD (Lee et al. 1985; Ayala 1986; Kwiatkowski et al. 1991). However, despite this correspondence with the covarion simulations, the problem with interpreting real sequences remains that similar decreases are predicted too by gamma models like equation (1) (fig. 1). The best fits of equation (1) to the intermediate curves (with \( b = 3/4 \)) are associated with SSDs less than 0.014. These results reinforce the earlier conclusion from figure 1 that the analysis of sequence differences alone is insufficient to discriminate between the covarion and gamma processes.

Both processes constrain the apparent evolutionary rates of sequences. Variability in the covarion model is limited by the size of the covarion pool and by its relatively slow rate of exchange with the temporarily invariant class. The apparent rate of sequence divergence declines over time as substitutions occur only among the covarions and exchanges become less successful in switching them with unvaried and temporarily invariant positions. In contrast, sites are always variable in the gamma model, but only a fraction of them evolve at the maximum rate. The decreased apparent rate in this case is the result of substitutions accumulating in rapidly evolving positions first and in more slowly evolving ones later. In these analogous ways, both processes lead to apparent rate slowdowns like those for SOD (fig. 1) and the intermediate curves in figure 3.

Estimates of \%Varied for \( c = 280 \) and 760 decline in the covarion simulations of II from their highs at \( v = 0.0 \) to their limits of covarion size (0.18 and 0.50 of \( n \)) at \( v = 1.0 \) (fig. 4A). Greater restrictions are imposed on the variability of sites as \( v \) increases, thereby limiting their effective numbers. An analogous pattern is observed for the gamma model with decreasing \( a \) (fig. 4B). Rate variation among sites rises as \( a \) decreases, thereby duplicating the same effect as large values of \( v \). \%Varied for the gamma process approaches that for the one-parameter equation with increasing \( a \). In contrast, a lower asymptote is approached in the covarion simulations, since 0.22 of their positions are permanently invariable and unable to change. Restrictions on variability occur in the covarion and gamma models even when \( v = 0.0 \) and \( a = 5.00 \), respectively. \%Varied under these conditions remains less than that for the one-parameter process where all positions are free to vary at the same rate (Jukes and Cantor 1969).

Estimates of \%FMP for \( c = 280 \) in the covarion hypothesis are lower than those for \( c = 760 \) and for the other models except when \( v = 1.0 \) (fig. 4C). Unless the exchange rate is very low (\( \sim 0 \)), the identities of covarions for the two groups diverge to where their smaller variable classes (0.18 of \( n \)) show reduced agreement as reflected by their lower \%FMPs. The same process occurs when \( c = 760 \) and the effects are dampened by the greater opportunities for overlap between their larger covarion classes (0.50 of \( n \)). Nevertheless, values of \%FMP for \( c = 760 \) also become large as \( v \) nears 1.0 (fig. 4A). Low rates of exchange constrain the divergence of covarion classes, thereby limiting the availability of temporarily invariable sites for change and causing the accumulation of multiple substitutions at the same covarion positions. The minimum values of \%FMP for \( c = 280 \) and 760 (at \( v = 0.7 \) and 0.6, respectively) represent the points where the two groups differ the most in the memberships of their covarion pools. In contrast, all \%FMPs for the gamma model exceed that for the one-parameter equation, which is approached as \( a \) increases (fig. 4B). This decline in \%FMP is the result of the rate variation among sites becoming more even (Law and Kelton 1991; Nei 1991).

SOD Fit by Models (III)

The \%FMP estimates in simulations II support the proposal that the varied and unvaried positions of two
distinct groups are more often the same in the gamma process than in the covarion hypothesis (fig. 4). These simulations are consistent with the original argument that the covarion and gamma models can be discriminated in this way, particularly when $c$ and $v$ for the former and $a$ for the latter are small, as for the mammal and plant SODs (0.18 of $n$, 0.01 [Fitch and Ayala 1994a], and 0.27 or 0.55 [fig. 1], respectively). This approach is followed in the tests of the covarion and gamma models in simulations III.

The first set of gamma-corrected branch lengths in figure 2 is based on equation (2) with $a = 0.27$, its optimal value according to the best fit of formula (1) to observed SOD differences (table 1 and fig. 1). The expected counts of the covarion hypothesis, given these corrected lengths and the other conditions in III, conform closely to the observed frequencies for mammal and plant SODs ($\chi^2_{13} = 2.51, P > 0.250$). Like the SOD sequences themselves, these predictions are characterized by increased numbers of FMPs and FmPs to FMPs as reflected by their %FMP estimates of 0.31 and 0.30, respectively. In contrast, the expectations of the gamma model with $a = 0.27$ differ significantly from the observed frequencies ($\chi^2_{13} = 22.86, P < 0.001$; table 1). These predictions over- and underestimate the numbers of FMPs to FMps and FmPs as indicated by their %FMP of 0.55.

A better fit to the observed SOD frequencies is obtained for the gamma process with its value of $a$ optimized to 0.55 (table 1). Indeed, its expectations with $a = 0.55$ become insignificantly different from the actual counts, although not by much ($\chi^2_{13} = 7.36, P > 0.050$). A closer fit to the observed data is also achieved by the covarion hypothesis when branch lengths are gamma corrected with this value of $a$ ($\chi^2_{13} = 1.01, P > 0.750$; table 1). Thus, the covarion model makes predictions that agree closely with the actual frequencies for mammal and plant SODs regardless of which set of gamma-corrected branch lengths is used (table 1 and fig. 2). This agreement stands in contrast to the poorer fits by the gamma model even when its value of $a$ is optimized to the SOD frequencies themselves.

Use of $a = 0.55$ weakens the fit of gamma equation (1) to observed SOD differences from $SSD = 0.009$ for $a = 0.27$ to 0.018 (fig. 1). This fit is worse than that for the unoptimized parameters of the covarion hypothesis ($SSD = 0.015$). Gamma equation (1) with $a = 0.55$ underestimates SOD differences early in sequence evolution but overestimates them later on. Acceptance of this value of $a$ carries with it an associated cost for predicting SOD divergence.

The predictions of the one-parameter process deviate the most from the observed frequencies for mammal and plant SODs ($\chi^2_{13} = 133.91$ and 90.49 for gamma-corrected branch lengths with $a = 0.27$ and 0.55, respectively, $P < 10^{-21}$ for each; table 1). The one-parameter equation overestimates the total number of hit positions as reflected by its %Varied estimates of 0.83 and 0.80, respectively. These values are much higher than those for the covarion (0.52–0.55) and gamma
Table 1
Observed and Expected Frequencies of Varied and Unhit Codons between Mammals and Plants and Their Simulated Representatives, Respectively

<table>
<thead>
<tr>
<th>Results</th>
<th>% Varied*</th>
<th>% FMP</th>
<th>Fmp</th>
<th>FMp</th>
<th>FMp</th>
<th>FMp</th>
<th>( \chi^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>0.49</td>
<td>0.31</td>
<td>78</td>
<td>24</td>
<td>27</td>
<td>23</td>
<td>...</td>
</tr>
</tbody>
</table>

Branch lengths, gamma corrected \((\alpha = 0.27)\):
- One-parameter model
  - Expected: 0.83 0.42 25.3 35.2 38.0 53.5 133.91
  - Observed: 0.55 0.30 68.6 39.1 29.2 25.0 2.51
  - (0.55) (0.30) (67.8) (28.9) (29.2) (26.2) (2.91)
- Covarion model
  - Expected: 0.80 0.38 31.0 36.3 38.9 45.3 90.49
  - Observed: 0.52 0.27 73.6 26.9 30.1 21.4 1.01
  - (0.53) (0.28) (72.2) (27.3) (30.1) (22.5) (1.20)
- Gamma model
  - Expected: 0.80 0.46 72.6 20.4 22.1 36.9 7.36
  - Observed: 0.52 0.46 72.6 20.4 22.1 36.9 7.36

Expected:
- Branch lengths, gamma corrected \((\alpha = 0.55)\):
  - One-parameter model
    - Expected: 0.55 0.30 68.6 39.1 29.2 25.0 2.51
    - Observed: 0.55 0.30 68.6 39.1 29.2 25.0 2.51
  - Covarion model
    - Expected: 0.55 0.30 68.6 39.1 29.2 25.0 2.51
    - Observed: 0.55 0.30 68.6 39.1 29.2 25.0 2.51
  - Gamma model
    - Expected: 0.55 0.30 68.6 39.1 29.2 25.0 2.51
    - Observed: 0.55 0.30 68.6 39.1 29.2 25.0 2.51

* Column headings are as follows: % Varied, \((\text{FMP} + \text{FMp} + \text{FmP})/\text{total sequence length}\); % FMP, \(\text{FMP}/(\text{FMP} + \text{FMp} + \text{FmP})\); Fmp, frequency of positions unhit in both mammals and plants; FMp, frequency of codons changed in mammals but not plants; FmP, frequency of positions varied in plants but not mammals; and FMP, frequency of codons varied in both taxa.

**Note:** Expected counts, which are each averages of 100 runs in simulations III, are compared to observed values by the \(\chi^2\) test with 3 degrees of freedom \((\chi^2_{36})\). These simulations are based on \(n = 152\) and the accepted phylogeny with both sets of its gamma-corrected branch lengths \((\alpha = 0.27\) and 0.55; fig. 2). Other variables in the gamma and one-parameter trials are set as in simulations II (fig. 4). Additional parameters in the covarion runs follow Fitch and Ayala (1994a), including their estimate of \(\alpha = 2.5\). However, this value of \(\alpha\) is not critical as very similar results for the covarion model are obtained when \(\alpha = 4\) (in parentheses).

Spatial Relationships of Replacements

The most variable regions of Cu, Zn SOD belong to \(\beta\)-strand hairpins and random coils that are located on the surface of the protein as a patch (Getzoff et al. 1989; Kwiatowski et al. 1992). These solvent-exposed regions of SOD correspond to residues 23–41 and 87–112 in the mammal and plant alignment. Estimates of %Varied and %FMP for these 45 positions are 0.78 and 0.49, respectively. These large values highlight the rapidly evolving nature of these codons, which share no apparent function other than contributing to the structural backbone of the protein. In contrast, the 21 amino acids of the active-site channel are highly conserved (table 2) because of their functional importance in catalysis (Tainer et al. 1983). Unlike the rapidly evolving group of surface residues, only 3 of the 21 positions for this feature differ in mammals and plants (%Varied = 0.14), with none of them changed in both taxa.

The active-site channel and \(\beta\)-barrel are the two outstanding features of tertiary structure for bovine SOD (Getzoff et al. 1983; Tainer et al. 1983). The frequencies of varied and unvaried codons between mammals and plants differ significantly among these structures and all other regions of the protein \((\chi^2_{16} = 18.45\) with 6 degrees of freedom, \(P < 0.010\); table 2). The largest differences are due to the uneven rates of change between the highly conserved active-site channel versus the more rapidly evolving \(\beta\)-barrel and all other positions. However, the unequal distribution of varied codons unique to mammals makes important contributions, too, to the \(\chi^2\) statistic. In contrast to their even distribution in plants, codons unique to mammals are under- and overrepresented in the \(\beta\)-barrel and all other regions, respectively. These results indicate that mammal and plant variabilities are distributed differently on the tertiary structure of SOD (Margoliash et al. 1972).

The first eight changed codons at the amino end of Cu, Zn SOD are varied only in plants (aligned positions 2, 4, 6, and 9–13). Codon 2 is part of the N-terminus, whereas residues 4, 6, and 9 and 10–13 belong to the first \(\beta\)-strand and loop of SOD, respectively (Tainer et al. 1983). The total number of changed positions in mammals and/or plants is 74. The probability of a string of eight or more consecutive varied residues, unique to either mammals or plants and anywhere along the sequence \((Pr)\), can be estimated by the equation:

\[
Pr = \sum_{i=8}^{j} \left(\frac{j}{74}\right)^{j-i+1},
\]

where \(j\) is the frequency of varied codons limited to plants \((\text{FmP} = 27)\) or mammals \((\text{FMp} = 24)\), and \(i\) is the length of all runs from 8 to \(j\) (table 2). Equation (5) provides a conservative, first-order approximation to the problem. The chance of a run of eight or more consec-
Distributions of Varied and Unvaried Codons in Mammals and Plants among the Active-Site Channel, β-Barrel, and All Other Regions of Cu, Zn SOD

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Active-Site Channel</th>
<th>β-Barrel</th>
<th>All Other Regions</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positions in mammal/plant alignment†</td>
<td>45, 47, 57, 59-</td>
<td>4-9, 15-21, 28-35,</td>
<td>1-3, 10-14, 22-27, 36-39,</td>
<td>1-152</td>
</tr>
<tr>
<td></td>
<td>62, 64, 70, 79,</td>
<td>40-44, 46, 81,</td>
<td>48-56, 58, 63, 65-69,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>82, 119, 123,</td>
<td>83-88, 94-100,</td>
<td>71-78, 80, 89-93, 101-</td>
<td></td>
</tr>
<tr>
<td></td>
<td>132, 135-140,</td>
<td>114-118, 145-</td>
<td>113, 120-122, 124-131,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>142</td>
<td>149</td>
<td>133, 134, 141, 143,</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>144, 150-152</td>
<td></td>
</tr>
<tr>
<td>Observed and (expected) frequencies‡:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fmp</td>
<td>18 (10.8)</td>
<td>25 (26.1)</td>
<td>35 (41.1)</td>
<td>78 (78.0)</td>
</tr>
<tr>
<td>FMp</td>
<td>1 (3.3)</td>
<td>4 (8.1)</td>
<td>19 (12.6)</td>
<td>24 (24.0)</td>
</tr>
<tr>
<td>FmP</td>
<td>2 (3.7)</td>
<td>11 (9.1)</td>
<td>14 (14.2)</td>
<td>27 (27.0)</td>
</tr>
<tr>
<td>FMP</td>
<td>0 (3.2)</td>
<td>11 (7.7)</td>
<td>12 (12.1)</td>
<td>23 (23.0)</td>
</tr>
<tr>
<td>Total</td>
<td>21 (21.0)</td>
<td>51 (51.0)</td>
<td>80 (80.0)</td>
<td>152 (152.0)</td>
</tr>
</tbody>
</table>

Summary statistics§:

<table>
<thead>
<tr>
<th></th>
<th>% Varied</th>
<th>% FMp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed and (expected)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fmp</td>
<td>0.14</td>
<td>0.00</td>
</tr>
<tr>
<td>FMp</td>
<td>0.51</td>
<td>0.42</td>
</tr>
<tr>
<td>FmP</td>
<td>0.56</td>
<td>0.27</td>
</tr>
<tr>
<td>Total</td>
<td>0.49</td>
<td>0.31</td>
</tr>
</tbody>
</table>

† Positions in the mammal/plant alignment included among the three structural categories of SOD (Getzoff et al. 1983; Tainer et al. 1983). Four of the seven metal-liganding residues of the protein (marked with asterisks) are assigned to the active-site channel, even though they are part of the β-barrel too.

‡ Observed and (expected assuming all sites are equally variable) frequencies of varied and unvaried codons between mammals and plants. Abbreviations: FMP, frequency of codons changed in both mammals and plants; FMp, frequency of positions varied in mammals but not plants; FmP, frequency of codons changed in plants but not mammals; and Fmp, frequency of positions unvaried in both.

§ Summary statistics for the observed frequencies: % Varied = (FMP + FMp + FmP)/total sequence length; and % FMP = FMP/(FMP + FMp + FmP).

Conclusions

We conclude that the variable and invariable codons of mammal and plant SODs are different as predicted by the covarion model. These conclusions are supported by the close fits of the expected results for the covarion model to the observed frequencies and differences for mammal and plant SODs (particularly with regard to its prediction of reduced FMPs relative to FMps and FmPs) and by the different distributions of mammal and plant variabilities on the three-dimensional structure of the protein (tables 1 and 2 and fig. 1). These results are inconsistent with the expectations of the gamma model in which a site's variability does not change with the taxonomic group. Apparent rate slowdowns as for SOD (fig. 1) are attributed instead to temporal fluctuations in codon variability as currently variable and temporarily invariable positions are exchanged. The covarion process is capable of explaining the main features of SOD evolution, and its success in this study reinforces the need to consider the importance of rate changes among positions over time.

The equations for the one-parameter and gamma version of the one-parameter models in this study are based on the assumption that substitution rates are the same for all types of change. However, more complex equations are available which account for variation in the transformation probabilities of different substitutions (e.g., the well-known two-parameter formula of Kimura [1980] and its gamma version for rate heterogeneity among sites [Jin and Nei 1990]). For the gamma versions of the one-parameter and higher-parameter models, Rzhetsky and Nei (1994) recently developed new equations that do not seriously overestimate numbers of substitutions when divergence is high and/or sequences are short. Furthermore, all of these equations can be modified to accommodate the possibility of permanently invariable sites. For example, the gamma version of the one-parameter model (1) can be modified in this way as

\[ p = bx\{1 - [a/(a+2\lambda t/b)]^e\}, \]

where \( x \) is the proportion of potentially variable sites (i.e., \( n-pi \)), as in equation (4) (Palumbi 1989).

Other proteins and nucleic acids show apparent rate slowdowns similar to that for SOD (e.g., mitochondrial
DNA [Fitch 1986a, 1986b] and globins [Shoemaker and Fitch 1989]). The availability of these data and other approaches will permit valuable tests about the generality of the covarion hypothesis and the conclusions of our study.

Acknowledgments

The following individuals and institutions are thanked for their assistance and support. N. Takezaki provided the source code for his program GENSEQ with the encouragement of M. Nei. B. Serog and H. Van offered instruction and encouragement on the computer and mathematical analyses. S. Yannone in the laboratory of B. Burgess assisted with the investigation of SOD tertiary structure. M. Antezana, B. Serog, M. R. Tennant, H. Van, and two anonymous referees added improvements to the manuscript. This research was supported by grants from the National Science Foundation to M.M.M. (BSR-8857264) and W.M.F. (BSR-9096052) and from the National Institutes of Health to W.M.F. (GM42397).

LITERATURE CITED


SHOEMAKER, J. S., and W. M. FITCH. 1989. Evidence from nuclear sequences that invariable sites should be considered when sequence divergence is calculated. Mol. Biol. Evol. 6:270-289.


PAUL M. SHARP, reviewing editor

Received July 22, 1994

Accepted November 30, 1994