J. Mol. Evol. 12, 197–218 (1979)

Journal of Molecular Evolution

The Covarion Model for the Evolution of Proteins: Parameter Estimates and Comparison with Holmquist, Cantor, and Jukes' Stochastic Model

John M. Karon*

Department of Mathematics, The Colorado College, Colorado Springs, Co 80903, USA

Summary. W. Fitch used a mathematical model to estimate the covarion size (the number of codons which are variable at a given time) and the turnover rate of covarions in the evolution of cytochrome c. We improve and correct the mathematical derivations and statistical estimation procedures in Fitch's model, altered to account more fully for the redundancy in the genetic code. We also consider a closely related model, which assumes the covarion fixing the last minimum mutation distance increasing (MDI) substitution has the same probability of losing variability as the other covarions. The average number of covarions is estimated to be at most five. Roughly 35 to 65% of the covarions are predicted to lose variability after each MDI substitution; this is smaller than Fitch's estimate, but the estimate is quite sensitive to changes in the data, which are a phylogenetic tree derived by Fitch and Margoliash. Both covarion models predict that there are about 0.9 to 2.0 total substitutions per variable codon in cytochrome c during "short" periods of evolutionary time (at most 10 MDI substitutions). This is less than the prediction from Holmquist, Cantor, and Jukes' stochastic model, which emphasizes variability over the entire time of divergence, rather than variability at a given time as in the covarion model, but this difference is predicted by the differing model assumptions.

Both the covarion and interactive models provide clear descriptions for hypotheses of a stochastic evolutionary process operating within deterministic selective constraints. Both depend on only two parameters, one measuring selective constraints, and the other, the rate of a stochastic process. Both factors are important, so it is unlikely that one could describe the process with fewer parameters. Since both models provide similar estimates for the rate of substitution for closely related pairs of species, it is plausible that both describe the same process, but from different viewpoints. Extensive tests on the protein data using an improved covarion model are necessary to determine whether these models are in fact compatible.

* Current address: Department of Biostatistics, University of North Carolina, Chapel Hill, North Carolina 27514, USA

0022-2844/79/0012/0197/$ 04.40

Key-Words: Molecular evolution - Evolutionary rates - Cytochrome c - Codon variability - Mathematical modeling

Introduction

Two types of stochastic models for the evolution of proteins have been proposed to estimate the intensity of the selective constraints on the genes coding for proteins. Holmquist and his colleagues have obtained estimates of the total number of codons ever open to a substitution and of the rate of substitutions over the entire time of divergence between two species for a number of proteins. For a review of their interactive model based on random evolutionary hits, see Holmquist (1976) and references therein; and for the data, Jukes and Holmquist (1972), Moore et al. (1976), and Holmquist et al. (1976). This model provides information on the more significant effects of natural selection, the restrictions on how many and which particular residues in a protein can vary (Holmquist, 1976, p. 105).

In trying to answer this question, Fitch (1971) took an alternative point of view, appropriate for the relatively short time corresponding to a few observable nucleotide substitutions. He used a mathematical model for the covarion hypothesis of Fitch and Markowitz (1970) to estimate the number of codons open to a substitution at a given time and the rate at which these codons lose variability. The covarion hypothesis states that, at a given time, a nucleotide substitution in a gene coding for a protein can occur only in a specific subset, called the *covarions*, of the codons for the protein; a mutation at any other site would be lethal or sufficiently deleterious to be discarded by selective forces. Fitch (1971) showed it to be a necessary consequence of this hypothesis that the variable codons change with time. This is supported by the interactive model prediction that, in general, the number of potentially variable codons increases with (sidereal) time (Jukes and Holmquist, 1972).

A third model for protein evolution is the maximum parsimony evolutionary tree, constructed from a (presumed known) evolutionary tree, contemporary protein sequences, and an algorithm for reconstructing ancestral sequences so that the total number of substitutions in the tree is minimized (see Moore et al., 1976, and references therein). In contrast to the interactive and covarion models, this model is completely deterministic. The validity of both the maximum parsimony and Holmquist's random evolutionary hit stochastic model does not depend on the time span modeled, but the assumptions on which the covarion model is based are only likely to be reasonable approximations to reality for relatively short evolutionary times. The covarion model could be considered to be intermediate between the interactive and maximum parsimony models, as it has the stochastic character of the former but with stronger explicit selective constrainsts (e.g. the restrictions against back substitutions and codons' regaining variability; see the following section). Moore et al. (1976) have shown that the estimates from the interactive random evolutionary hit stochastic and maximum parsimony models are consistent, particularly for long evolutionary times. I will provide preliminary evidence that the interactive and covarion stochastic models are consistent for relatively short evolutionary time periods.

Biological knowledge supporting the covarion hypothesis includes the constraints on the set of potentially variable codons in cytochrome c based on functional requirements. In addition, Fitch and Markowitz (1970) discussed the spatial correlation in

amino acid replacements, suggesting that a potentially variable codon can lose its variability as a result of a substitution at another codon (or perhaps a change in another protein with which cytochrome c interacts). Further evidence for loss in variability is provided by the (perhaps nearly irreversible) substitution of alanine for cystine at residue 14 in the cytochrome c of *Euglena gracilis* (Pettigrew 1973 and Lin et al. 1973).

The discussion of the covarion model shows that the simplest parameters required to describe it are the average number of covarions and the rate at which codons lose variability. Fitch estimated these parameters for cytochrome c by computing the best fit of the values predicted by his mathematical model to the minimum mutation distances in the phylogenetic tree of Fitch and Margoliash (1968). I have corrected and improved Fitch's mathematical and statistical methods; see Appendix A and Methods for details. My estimates confirm his prediction that the number of covarions is small (five or less)[1], but suggest that 35 to 65% of the covarions lose variability after each nucleotide substitution changing the minimum mutation distance, in contrast to his prediction of at least 60%.[2] As pointed out in Results, this difference is to be expected as a result of greater use of the redundancy in the genetic code in my models than in Fitch's.

The estimates given here should be regarded as qualitative, rather than quantitative, especially that for the turnover rate of covarions, which is quite sensitive to small perturbations in the data. (The estimate of the covarion size is affected very little by these perturbations.) More complete data and a slightly better model, as outlined in the final section, are necessary before good quantitative estimates can be made.

Unless specifically stated to the contrary, all references to Fitch's work are to Fitch (1971).

Two Models for the Covarion Hypothesis

I will analyze two models, each of which describes a series of nucleotide substitution events. It will be necessary to differentiate between nucleotide substitutions which increase the minimum mutation distance (MMD) between the encoded amino acids, and those which do not; the former will be called mutation distance increasing (MDI) substitutions.

In addition to the description of the covarion hypothesis in the Introduction, the models are based on the assumptions listed below. These particular assumptions are made in order to obtain models which are as simple as possible while reproducing important features of the evolutionary process. Examples contrary to assumptions 2-5, describing the nucleotide substitutions allowed by the model, are known, but each is likely to be more reasonable as the period of evolutionary time modeled is decreased.

---

1 This estimate also agrees qualitatively with Fitch's latest estimate (Fitch 1976), based on a linear extrapolation, of a covarion size in plants and mammals of about 10 to 12 codons. The data plot suggests that a quadratic extrapolation might provide a somewhat better fit and a lower estimate of covarion size.

2 This is one minus the expected proportion of covarions remaining variable after a substitution (defined in Results), computed from Fitch's estimate of about five covarions and a probability of retaining variability of at most 0.25.

Since the data fit will be restricted to 10-15 MDI nucleotide substitutions, it is only necessary that these assumptions be true over the time required for about 10 MDI substitutions. Fitch made the same assumptions, except that he allowed only MDI substitutions in assumption 4 and considered only Model I in assumption 6.

Assumption 1. The covarion hypothesis is valid, and at any time there are exactly c covarions, where c is an integer. The probability v that a covarion remains variable after a substitution (see Assumption 6 for a more precise definition) is also constant.

I know of no direct evidence on whether the parameters are roughly constant for reasonably closely related species, although Fitch (1976) has presented data suggesting that the number of covarions is about the same in mammals and plants. Allowing the basic parameters c and v to vary, or c to be nonintegral, would make the mathematical analysis quite complex. However, the estimates can be extended to nonintegral values of c (see Results). Both parameters should be regarded as representing average values over a number of evolutionary paths.

Assumption 2. No MDI substitutions occur in a codon which has lost its variability after fixing an MDI substitution.

This assumption may be reasonable during a sufficiently short period of time, particularly if substitutions are selectively advantageous. It greatly simplifies the computations, as a random process would have to be chosen (and perhaps another parameter introduced) to describe codons regaining variability after fixing an MDI substitution. This assumption is used in the computations summarized in Results. In the discussion of methods and the comparison with the REH interactive model, the stricter assumption (*Assumption 2a*) that no codon can regain variability will be made in order to simplify some computations.

Assumption 3. Back substitutions (i.e. substitutions decreasing the MMD) do not occur.

We must make this assumption to fit MMD data in a phylogenetic tree, since the data are computed under the hypothesis that back substitutions do not occur on an internodal leg of the tree. This does not appear to be a severe restriction for random mutations. For codons with MMD=1, back substitutions occur with a frequency of only 24/1500 through random mutations (see the derivation of the proportion of second MDI hits on codons with MMD=1 yielding MMD=2, which will be denoted by u, later in this section). Further data on random back substitutions are the below diagonal entries in Table A2 in Holmquist et al. (1972). In 230 substitutions in a phylogenetic tree for cytochrome c, Fitch and Margoliash (1968) found only three back mutations; in their reconstruction of an ancestral myoglobin sequence using an MMD cladogram, Romero–Herrera et al. (1973) found only seven detectable back substitutions in the 147 substitutions required.

Assumption 4. All base substitutions which increase or do not change the MMD are equally likely at each base position in each covarion with an MMD of zero or one, and each such covarion is equally likely to fix the next such substitution. (In addition, see the next assumption for the treatment of codons with MMD=2).

This assumption was made to obtain a mathematically tractable model. Specific probabilities for different types of base changes could be incorporated in the model by altering the method used to compute the parameter u.

Assumption 5. Covarions with MMD=2 are treated as if they had MMD=1 in the substitution process, except that the possibility of a codon with MMD=3 is ignored in the calculations.

Although pairs of species with MMD=3 for cytochrome c are known, these are quite rare (Moore et al., 1976, p. 27), and there are none among those pairs with MMD at most 15 in the 57 sequences tabulated by Holmquist (personal communication; this tabulation is the one from which Tables 2-5 in Moore et al. (1976) were derived). The first part of this assumption is not a serious restriction, as covarions with MMD=2 are unlikely to occur very often when the total number of substitutions is small. This hypothesis is used in computing the probability p_k in Equation (1), below, which specifies the probability that the next MDI substitution will occur in a covarion with MMD=0. Thus, after two MDI substitutions p_k should be slightly larger than the value given, but the difference between the correct value and that in (1) is small.

Assumption 6. After each MDI substitution, the covarions are subject to a loss of variability according to a binomial random process:

a.) For Model I, the covarion which fixed this mutation remains variable; each of the other covarions remains variable with probability v (hence loses variability with probability 1-v).

b.) For Model II, each covarion remains variable with probability v. Hence, the codon fixing the last MDI substitution may also lose variability.

MDI substitutions initiate the replacement process since the data to be fit are MMD's from a phylogenetic tree. In the absence of any information on how the set of covarions changes with time, the assumption that loss of variability occurs independently and with the same probability in each codon (Model II), or in a particular subset (Model I), seems to me to be the simplest. One of the motivations for Fitch and Markowitz's originally proposing the covarion hypothesis was the unexpectedly high proportion of codons with MMD=2 in the phylogenetic tree for cytochrome c. Fitch gives data supporting the inherent bias in Model I, which is his model, toward double substitutions (the bias results from the codon fixing the last substitution remaining variable). Since it is conceivable that this codon could lose variability, Model II will also be considered. In fact, it seems reasonable that the codon fixing the last substitution may lose variability, but with a smaller probability than the other covarions, e.g. if an optimal substitution requires two substitutions instead of just one (see Fitch's discussion). Thus the predictions from these two models may be regarded as bounds for the true turnover rate, and comparing them would indicate whether the difference is important.

The first two assumptions imply that the potentially variable codons (those not absolutely fixed in order that the protein function) can be divided into three classes: those which have never been variable, or which did not fix an MDI substitution while variable (group I); the covarions (group II), of which there are always c, an integer; and the codons which were once variable but are no longer and have fixed an MDI substitution (group III). Thus, if the gene has N potentially variable codons, initially

there are N-c codons in group I, c in group II, and none in group III. After each MDI substitution, codons are removed from group II (and enter group III or return to group I) according to the binomial random process specified above, with turnover rate 1-v. Codons from group I then become variable, to maintain c codons in group II.

Explicitly, each MDI nucleotide substitution in the gene consists of the following sequence of events:

1. *Gene substitution*. Let there be k covarions with an MMD of zero. The next MDI substitution is fixed in one of the previously unhit covarions with probability p_k (see Equation (1), below), and in one of the previously hit covarions (MMD=1) with probability $1-p_k$. If a previously unhit (hit) covarion fixes the substitution, all k (respectively c-k) such covarions are equally likely to fix it.
2. *Loss of variability*. For model I, the covarion which fixed this MDI substitution must remain variable; for Model II, it need not. Each covarion subject to loss of variability (c-1 codons in Model I, c in Model II) remains variable with probability v and is transferred to group II with probability 1-v.
3. *Replacement of lost covarions*. If a total of j ($0 \leqslant j \leqslant c$) covarions reenter group I and are transferred to group III, j codons are transferred from group I to group II.

To complete the specification of the model, we need only define the probability p_k, introduced above, that the next MDI substitution occurs in one of the k covarions with MMD=0. This probability is determined by the redundancy in the genetic code, for this redundancy implies that a nonsynonymous substitution — a substitution changing the encoded amino acid — in a codon with MMD=1 is less likely to increase the MMD than a substitution in an unhit codon. The same base could be hit twice (as AAA→CAA→GAA), or hits on two bases could yield an MMD of 1 (as GAA→GAU→GCU, Glu→Asp→Ala) or 0 (as UCU→ACU→AGU, Ser→Thr→Ser). Suppose that a codon fixes two non-synonymous mutations; Fitch (1971, p. 94) showed that the two hits are on different positions with probability 0.604. By direct enumeration (using a computer), we find that the number of such second substitutions yielding an MMD of 0, 1, 2 are, respectively, 24, 476, and 1000. Ignoring the first possibility, since it corresponds to a back substitution, the proportion of second MDI hits on codons with MMD = 1 yielding MMD = 2 is

$$u = 0.604 \times (1000/1476) = 0.416.$$

We can use this information to compute p_k. Let p be the probability of a nonsynonymous substitution (note that p is independent of MMD); k ($0 \leqslant k \leqslant c$), the number of covarions with MMD=0; and π_i, the probability that the next substitution is an MDI substitution in some covarion with MMD = i, i = 0 or 1. Then $\pi_0 = p \cdot k/c$ and $\pi_1 = p \cdot u \cdot (c-k)/c$. Now observe that p_k is also the conditional probability that a substitution occurred in some covarion with MMD=0, given that it was an MDI substitution. Since a conditional probability is defined as a ratio of two probabilities, we have

$$p_k = \pi_0/(\pi_0 + \pi_1) = k/(k + u(c-k)) \quad k = 0,1,\ldots,c \qquad (1)$$

where u is the proportion of hits on covarions with MMD=1 resulting in MMD=2; its numerical value is given above. This is the probability used in step one of the model.

It is important to note that there are essential differences between protein evolution and these models. The models assume that evolution during an internodal interval on the phylogenetic tree is independent of past history and that, at the start of each substitution event, the future behavior of the model is independent of previous events, except for the number of covarions carrying one or two substitutions. Also, the models do not provide for parallelism, which Holmquist (1976) and Romero–Herrera et al. (1976) have discussed as a limitation of the maximum parsimony method.

The only difference between Model I and Fitch's model is the value of the parameter u. Fitch considered only MDI substitutions and did not attempt to correct fully for the redundancy in the genetic code; he used the value u=0.604, rather than 0.416. As a result, one would intuitively expect these models to predict a slower turnover rate and/or a smaller number of covarions than Fitch's model (I am indebted to Fitch for this observation); in fact, a slower turnover rate is predicted.

Appendix A

Computation of the Probability of no Double Substitution

The basic idea for computing the probability of no double substitution is the following. According to the model, no double substitution in m minimum mutation distance increasing (MDI) substitutions means that each MDI substitution must occur in a covarion with MMD=0. At each step, the probability of this happening depends on the number of covarions with MMD=0 (see the discussion of the probability p_k defined in Equation (1) in the definition of the model). Therefore, it is necessary to know the probability that there will be a given number of covarions with MMD=0 at the beginning of each substitution and codon replacement (model step). It is then possible to use these probabilities to compute the corresponding probabilities at the end of the next model step.

It is easiest to carry out this intuitive idea by interpreting the model as an absorbing Markov chain. As indicated above, we divide the possible outcomes after each model step into c+2 possible outcomes ("states") based on the number of covarions with MMD=0 after the *mth* MDI substitution and replacements; this number will be denoted by n_m. Suppose there have been j MDI substitution events. For i=0, 1,..., c, state i is: no covarion has attained MMD=2, and $n_j = i$. The only alternative is state c+1: some covarion has attained MMD=2.

Let

$\pi_i^{(m)}$ = Pr(state i after m MDI substitutions), i=0,1...,c+1

$\pi^{(m)}$ = row vector, of length c+2, of the $\pi_i^{(m)}$

p_{jk} = Pr (state k after m model steps | state j after m-1 model steps)

where Pr(E) is the probability of the event E, and Pr(...|...) is a conditional probability. The model implies that the "transition" probabilities p_{jk} are independent of m: they depend only on j and k. (This will be clear from the discussion of their computation.) Let P be the square matrix of size c+2 containing the probabilities p_{jk}, with j the row index. It follows from the theory of Markov chains (see e.g. Parzen, 1962) that

$$\pi^{(m)} = \pi^{(o)} \mathrm{p}^m \qquad (2)$$

where P^m is the matrix P raised to the m*th* power. The desired probability of no double substitution is

$$w_m(c,v) = 1 - \pi^{(m)}_{c+1} \tag{3}$$

Relation (2) can also be interpreted as $\pi^{(m+1)} = \pi^{(m)} P$, which means that one obtains the probabilities $\pi^{(m+1)}$ for the next step by multiplying the probabilities $\pi^{(m)}$ by the matrix P.

To compute the transition probabilities p_{jk}, first note that it is impossible to leave the state with MMD=2 (it is "absorbing"), so

$$p_{c+1,k} = \begin{cases} 1 & k = c+1 \\ 0 & k = 0, 1, \ldots, c. \end{cases} \tag{4}$$

In addition, the c+2 states exhaust all possibilities, so from any state j one must reach some other state. Therefore,

$$\sum_k p_{jk} = 1 \text{ for each } j \tag{5}$$

which is used to compute $p_{j,c+1}$ from p_{jk} for $k \leqslant c$. The latter probabilities can be computed by noting that, if we define E_m to be the event that the m*th* MDI substitution is at a covarion with MMD=0, then

$$p_{jk} = \Pr(E_m | n_{m-1}=j) \cdot \Pr(n_m=k | E_m \text{ and } n_{m-1}=j) \quad j, k = 0, 1, \ldots, c \tag{6}$$

since the MMD of the covarion fixing the MDI substitution is independent of the following replacement process. The first factor in (6) is precisely the probability p_j given in (1).

The second factor in (6) is determined by the replacement process. Given that the event E_m has occurred, the last MDI substitution was at a covarion with MMD=0, so there are j-1 covarions with MMD=0, all subject to the replacement process. If any of these lose variability, their replacement codons also have MMD=0. Therefore we must have at least j-1 covarions with MMD=0 after the replacement process ($p_{jk}=0$ if $k \leqslant j-2$), and n_m is j-1 plus the number of covarions with MMD=1 subject to the replacement process which are actually replaced. For model I, the covarion fixing the last substitution is not subject to replacement, so the transition from $n_{m-1}=j$ to $n_m=k$ requires that k-(j-1) of the c-j covarions with MMD=1 subject to replacement lose variability. With the binomial replacement model, the probability of this occurring is

$$\Pr(n_m=k | E_m \text{ and } n_{m-1}=j) = \begin{cases} \binom{c-j}{k-j+1} v^{c-k-1} (1-v)^{k-j+1} & j-1 \leqslant k \leqslant c-1 \\ 0 & k=c \end{cases} \tag{7}$$

For Model II, the covarion fixing the MDI substitution need not remain variable, so k-(j-1) of the c-j+1 covarions with MMD=1 must lose variability, and this has probability

$$\Pr(n_m=k | E_m \text{ and } n_{m-1}=j) = \binom{c-j+1}{k-j+1} v^{c-k} (1-v)^{k-j+1} \quad j-1 \leqslant k \leqslant c. \tag{8}$$

To summarize, the transition probabilities p_{jk} are computed using (4) - (8). Then $\pi^{(m)}$ is computed from (2) and the fact that

$$\pi_i^{(0)} = \begin{cases} 1 & i=c \\ 0 & \text{otherwise} \end{cases}$$

Finally, the probability of no double substitution in m steps is obtained from (3).

This procedure can be used for integral values of c, but these probabilities are also needed for nonintegral c to compute the means and standard errors in Tables 2 and 3. The latter probabilities were obtained using a quadratic interpolation process, as outlined near the end of the next section in the discussion of the computation of the optimal value of WS(c,v).

Fitch erred in computing conditional probabilities by substituting an expected value in the conditioning event. He did this in his equation (1), where he computed the probability that the i+1 *th* MDI substitution occurs in a particular codon with MMD=0 (call this event E_i), given that there are c-e_i codons with MMD=0 after i MDI substitutions. The correct procedure requires the computation of $Pr(E_i \mid k_i)$ for $k_i = 1,2,...,c$, where k_i is the number of codons with MMD=0 after i MDI substitutions. The probabilities $w_m(c,v)$ can be obtained from these conditional probabilities, using conditional probability methods and recursion relations, but the Markov chain approach is easier to explain.

Methods

The parameters c and v were estimated by finding those values which minimize a weighted sum of squares WS(c,v), similar to a Pearson chi-squared, of the deviations of the data from the model predictions; WS(c,v) is defined later in this section. Such a procedure is generally referred to as finding the "best fit" of the model predictions to the data. The data used were the number of times a double substitution was, and was not, found for given numbers of substitution events in a phylogenetic tree for the cytochrome c of 20 species (Fitch and Margoliash, 1968). For reasons given below, these data, summarized in Table 2, were restricted to $m \leqslant 15$ minimum mutation distance increasing (MDI) substitutions and were grouped in computing WS(c,v). A numerical procedure, described below, was used to minimize WS(c,v).

The Definition of WS

Before defining WS(c,v), we recall the definition of the Pearson chi-squared. Suppose that on each of N independent experiments there are k possible outcomes (these may be groups of observed outcomes). Let d_i be the observed number of outcomes of type i, $i = 1,2,...,k$, and let p_i be the model probability of observing such an outcome, so that Np_i is the number of outcomes of type i predicted by the model. Then the chi-squared is

$$\chi^2 = \sum_{i=1}^{k} (Np_i - d_i)^2 / Np_i \ ,$$

or the sum of (expected - observed)^2/expected. Note that χ^2 gives a weighted least squares estimate of the deviation of the model predictions from the data with weights $(Np_i)^{-1}$, so choosing the parameters in a model to minimize χ^2 results in a "best" fit

of the model to the data in this sense. In using the chi-squared, it may be necessary to group some of the outcomes so that most of the expectations Np_j are not too small (Lancaster 1969, p. 77). Intuitively, if Np_j is "small", the corresponding term in the sum is given too much weight.

To fit our data, first group the data into k cells according to the guidelines for using the chi-squared, each cell containing the data for one or more numbers of MDI substitutions. Let $Z_j(c,v)$ and z_j be, respectively, the total number of legs in the phylogenetic tree in cell j on which no double substitutions are expected and observed. The Z_j and z_j are computed using the probabilities $w_m(c,v)$ that no double substitutions occur in m substitution events given the parameter values c and v, derived in Appendix A, above, and the data in Table 2. For example, if the first cell contains the data for m=2 and m=3, there are six observations for m=2 and two for m=3, so $Z_1(c,v) = 6w_2(c,v) + 2w_3(c,v)$, and $z_1 = 7$. If cell j contains N_j observations and Y_j and y_j are, respectively, the number with at least one double substitution expected and observed, clearly $Y_j(c,v) = N_j - Z_j(c,v)$ and $y_j = N_j - z_j$. The weighted sum of squares analogous to the chi-squared is

$$WS(c,v) = \sum_{i=1}^{k} \left\{ (Z_j(c,v) - z_j)^2/Z_j(c,v) + (Y_j(c,v) - y_j)^2/Y_j(c,v) \right\}$$

for this is the sum of (expected - observed)^2/ expected over all possible outcomes.

However, WS is not the Pearson chi-squared. When the data are grouped into cells, each cell includes data for several values of the parameter m; for the chi-squared, a cell contains data for certain types of outcomes with all parameters fixed. In addition, the method for constructing the phylogenetic tree results in the data not being statistically independent. For these reasons, no significance levels will be given for the results.

Restrictions on the Data

Two considerations lead to using only the data for at most 10 to 15 MDI substitutions in the phylogenetic tree to fit the models. The data in the tree, which are inferred from sequence data, are made more reliable by this restriction. Second, the model assumptions, especially the restrictions against back substitutions and regaining variability, are likely to be reasonable only for a relatively short period of evolutionary time. It should be noted that we are discarding about one-third of the data in Fitch and Margoliash's phylogenetic tree.

In addition, let us temporarily make the stronger Assumption 2a, that no codon can regain variability. Then we can compute a lower bound for the number of MDI substitutions we could allow by requiring that the number of covarions which lose variability not exceed the number of potentially variable codons in the gene (if the latter occurred, codons must regain variability). For example, consider Model I. At each repetition of step 2, the number of sites transferred from group II to group III is random but has a binomial distribution: there are c-1 "trials" with a transfer probability of 1-v on each. There are m(c-1) such binomial trials after m substitutions, so the mean number of codons transferred is CT = m(c-1)(1-v). (For model II, CT = mc (1-v).) Thus, after m MDI substitutions under the modified assumptions there are c covarions in group II, an average of CT = m(c-1) (1-v) codons in group III, some potentially variable codons

in group I, as well as codons which have never been variable, e.g. fixed for structural or functional reasons. Let N be the number of potentially variable codons during the time period modeled. The above discussion shows that we must require $c + CT \leqslant N$. Since there were 25 sites which carry the same amino acid in all the cytochrome c sequences known in 1976 (Fitch 1976), a reasonable upper bound for the number of potentially variable codons may be $N = 80$. In searching for a best fit, values of c as large as 10 and of v smaller than 0.1 were used. This gives values of $c + CT$ as large as $10 + 8.1m$ for Model I, which implies $m \leqslant 9$. Since the assumption against regaining variability used to obtain this bound is too strong, this computation suggests that allowing up to about 15 MDI substitutions in fitting the model is appropriate.

Computation of Optimal Values of the Parameters

After the data were grouped into cells, the values of c and v for which WS (c,v) is a minimum were computed using the empirical observation that, for fixed c, WS(c,v) is virtually always concave up as a function of v. For each integer value of c between two and ten, the value of v minimizing WS was computed; denote this value by v*. Since v* changed quite rapidly with successive values of c, the optimal value of v (for all c) was then computed by finding the point on the parabola passing through the three points (c,v*,WS(c,v*)) in three-dimensional space for which WS was smallest; the location of this point also determines the optimal value of c. For example, consider Model I with the data from the phylogenetic tree grouped into three cells (see Table 2) and the parameter $u = 0.416$. For integral c, it was found that the smallest values of WS are WS(3,.00) = .74, WS(4,.16) = .39, and WS(5,.37) = .48; these can be interpreted as heights above the c-v plane. The smallest distance to the c-v plane from the parabola passing through these three points in three-space (the minimum value of WS on this parabola) is 0.37, and the corresponding optimal values of c and v are 4.30 and 0.22, respectively. Without using quadratic interpolation, one would expect the optimal values of c and v to be between four and five, and 0.16 and 0.37, respectively, but interpolation is required to obtain more precise estimates, especially for v.

Improvements in Fitch's Methods

This analysis contains several improvements over Fitch's work. Fitch used all of the data in the phylogenetic tree in fitting his model. Since he did not group the data into cells, he gave substantial weight to outcomes with relatively small expectations in computing the best fit, including the data for large numbers of MDI substitutions which do not fit the model. In addition, his use of a weighted sum of squares to fit the average number of codons with an MMD of two in the internodal intervals is less sensitive to perturbations in the data and the value of the parameter u, as well as less robust, than the quantity WS(c,v) proposed here. For model I, the estimates using Fitch's sum of squares (with data grouped into cells) are less robust because they are much more sensitive to the grouping of the data than are the estimates from WS. For model II, the estimates using his method are nearly independent of the changes in the parameter u and perturbations in the data considered in Tables 2 and 3, and are within the range of the estimates using WS.

Results

Given the simplifying assumptions made in these models and the small amount of data used to fit them, we can hope for at best qualitative results, rather than expecting parameter estimates of great accuracy. In addition, the values of the parameters c and v could be both time and species dependent, so the parameter estimates should be regarded as estimates of average values.

The best estimates of the parameters c and v, given in Table 1, suggest that, for Model I, v may be between .15 and .50, and c, between 4.0 and 4.5; for Model II, v may be between .35 and .60, and c, between 2.5 and 3.0. The nonintegral values of c arise from the computation of the best estimates of v, outlined in Methods. Cases 2-5 of Table 1 indicate the stability of these results under modification of the model assumptions and perturbations in the data, as discussed below. They show that the number of covarions is small, but the estimate of the turnover rate is sensitive to the model assumptions and the data.

Tables 2 and 3 contain comparisons between the model predictions and the original data, for fits of ungrouped and grouped data, respectively. The only estimates of the mean number of legs in the tree with no double substitutions, E(Z), greater than two standard errors from the observed number of legs with no substitution, z, are those for

Table 1. Best estimates of c and v for cytochrome c from the covarion model

Data case	u	z for m=10	Model I			Model II	
			c	v	EPRV	c	v
1	0.416	2	4.30	0.22	0.40	2.90	0.47
			4.14	0.16	0.36	2.64	0.36
2	0.300	2	3.48	0.18	0.42	2.49	0.25
			3.43	0.14	0.39	1.80	0.07
3	0.500	2	4.62	0.18	0.36	3.17	0.49
			4.44	0.07	0.28	2.99	0.43
4	0.416	1	4.70	0.55	0.65	3.02	0.62
			4.14	0.39	0.54	2.83	0.55
5	0.416	0	6.00	1.00	1.00	3.09	0.81
			5.12	0.81	0.85	3.00	0.77

The values of c and v are those minimizing the statistic WS(c,v), using the data from Table 2, modified as indicated in column 3 for the last two cases. WS(c,v) is defined in Methods, and z is the observed number of internodal legs on which a double substitution did not occur. EPRV is the expected proportion of covarions remaining variable after a substitution event. For each case, the estimates in the first line are for the data grouped into three cells (m=2,3; m=4,5; m $\geq$ 6), and in the second line, two cells (m=2,3,4; m $\geq$ 5), where m is the number of substitution events. The alternative values 0.30 and 0.50 for u were chosen arbitrarily, to bracket the value 0.416 computed in the definition of the model

Table 2. Number of legs on the phylogenetic tree with no double substitutions: Comparison between data and model predictions (mean ± one standard error), using best estimates of the parameters

Number of legs		6	2	4	2	1
Data: subst./leg (m)		2	3	4	5	6
No double subst. (z)		5	2	1	1	0
Predictions:		mean number with no double substitutions ± one standard error				
model, case	(c,v)					
II, 1	(2.90,0.47)	5.49 ± 0.68	1.59 ± 0.57	2.70 ± 0.94	1.14 ± 0.70	0.48 ± 0.50
I, 1	(4.30,0.22)	5.33 ± 0.77	1.52 ± 0.60	2.58 ± 0.96	1.09 ± 0.70	0.46 ± 0.50
II, 4	(3.02,0.62)	5.37 ± 0.75	1.45 ± 0.63	2.19 ± 1.00	0.81 ± 0.69	0.30 ± 0.46
I, 5	(5.12,0.81)	5.45 ± 0.71	1.48 ± 0.62	2.11 ± 1.00	0.66 ± 0.66	0.18 ± 0.39
Total no. legs		1	1	2	1	
Data:	m	7	9	10	15	
	z	0	0	2	0	
Predictions:		mean number with no double substitutions ± one standard error				
model, case	(c,v)					
II, 1	(2.90,0.47)	0.41 ± 0.49	0.29 ± 0.45	0.49 ± 0.61	0.10 ± 0.30	
I, 1	(4.30,0.22)	0.36 ± 0.49	0.28 ± 0.45	0.47 ± 0.60	0.10 ± 0.30	
II, 4	(3.02,0.62)	0.22 ± 0.41	0.12 ± 0.32	0.17 ± 0.40	0.02 ± 0.13	
I, 5	(5.12,0.81)	0.10 ± 0.29	0.02 ± 0.15	0.02 ± 0.16	0.00 ± 0.00	

The data are from the phylogenetic tree of Fitch and Markowitz (1968). m is the number of internodal substitution events; z, the number of legs in the tree with no double substitution. The mean is the expected value of Z, a random variable equal to the number of legs on which no double substitution occurs in the model. The values of the parameters c and v were chosen from those in Table 1, to which entries the case numbers refer. Corresponding comparisons for the data grouped into cells are given in Table 3. For each value of m, Z is a binomial random variable. Therefore, if the probability of no double substitution in m model steps is p, we have $E(Z) = np$ with a standard error of $(np(1-p))^{1/2}$ for n legs

m = 10 substitutions, and the data in Table 2 suggest that this particular datum may be a statistical outlier, i.e. an abnormally large deviation from its mean value.

To compare the results for these models, we must take into account the fact that one more covarion is subject to loss of variability in Model II than in Model I. The estimates of c agree with this difference: the estimate of c for Model II is 1.0 to 1.5 smaller than that for model I. To compare the turnover rates, we compute the expected proportion of codons remaining variable (EPRV) after each minimum distance increasing (MDI) substitution; this is also a more readily interpretable measure of turnover rate than v for Model I. Clearly this is v for Model II. For Model I, c-1 codons remain variable, each with probability v, as well as the codon fixing the last MDI substitution. Thus the mean number remaining variable is 1+v(c-1) of the c codons, so for Model I

Table 3. Number of legs in the phylogenetic tree with no double substitution: Comparison between grouped data and model predictions

	Group number	1	2	3
Data:	Total no. legs	8	6	6
	No double subst. (z)	7	2	2
Predictions:		mean number with no double substitution ± one standard error		
Model	(c,v)			
II	(2.90,0.47)	7.08 ± 0.89	3.84 ± 1.17	1.77 ± 1.08
I	(4.30,0.22)	6.85 ± 0.98	3.67 ± 1.19	1.70 ± 1.07

	Group number	1	2
Data:	total no. legs	12	8
	No. double subst. (z)	8	3
Predictions:		mean number with no double substitution ± one standard error	
Model	(c,v)		
II	(2.64,0.36)	10.15 ± 1.22	3.74 ± 1.35
I	(4.14,0.16)	9.40 ± 1.38	2.87 ± 1.29

The data is the same as in Table 1. The number of legs in the tree on which no double substitution occurs is z. The values of the parameters c and v are those from data case 1 of Table 1. For the grouped data, with three groups the groups are m = 2, 3; 4, 5; and 6 or more. With two groups, the groups are m = 2, 3, 4; and 5 or more

EPRV is $(1+v(c-1)/c = v+(1-v)/c$. The values of EPRV given in Table 1 suggest that it may be about 50% (roughly, between 35% and 65%). On the basis of direct comparisons of sequences, Jukes and Holmquist (1972) suggest that there is a high rate of change in the variable codons in cytochrome c, which would agree qualitatively with the conclusion that c is small but EPRV is not: such a combination would make it likely that a codon would accumulate a number of substitutions while it was variable.

The sensitivity of such estimates to perturbations in the data is always an important question, especially in a case such as this where the data are inferred from measured data. The data for m=10 may represent a statistical outlier, and the discussion of the previous section suggests that we would prefer that our estimates not be influenced too strongly by the data for such a large number of MDI substitutions. Therefore, the best estimates of c and v were also computed for perturbations in the data for m=10 (cases 4 and 5, Table 1). Since the estimates of EPRV are quite sensitive to these perturbations, the actual value of EPRV is rather uncertain.

One would also like to evaluate the sensitivity of these estimates to the assumptions on which the models are based, particularly the restrictions against back substitutions and condons' regaining variability. Ignoring back substitutions is not likely to be as severe for codons with MMD=1 as for codons with MMD=2 (see the discussion of the model assumptions). If the number of potentially variable codons is large, the restriction against regaining variability is not likely to be serious for 5-10 model steps;

however, it seems difficult to determine the severity of this restriction. It would be necessary to compare the numbers of potentially variable codons (i.e. the number which could become variable during the next replacement process) which have and have not previously been variable. As Table 5 shows, the interactive model predicts that the number of potentially variable codons is fairly small for at most 10 MDI substitutions; this is evidence that the restriction against regaining variability may be important.

The parameter estimates corresponding to the perturbed values of the parameter u (cases 2 and 3, Table 1) suggest the dependence of our results on the assumptions listed above. Its specification in the definition of the models shows that an increase in u increases the probability that a hit on a codon with MMD=1 will increase the MMD to two. Since allowing codons to regain variability would also increase this probability, the results for u=0.50 suggest the influence of this assumption. Similarly, decreasing u decreases this probability and corresponds to allowing back substitutions and hits on codons with MMD=2. While it is difficult to estimate the combined effect of all three assumptions, it is clear that the stability results are the same as those for perturbations in the data.

It can be seen from Table 1 that the predicted value of v for Model I decreases as the parameter u increases, while the estimated value of c increases slightly. This suggests that Fitch may have been substantially correct in his estimate of v as .25 or less for his model (Model I, with u = 0.604), despite the error in his probabilistic derivations (see Appendix A); however, the sensitivity of the estimates to perturbations in the data also makes this estimate uncertain. As Fitch (personal communication) has commented, "If one takes into account fixations that are not observable but whose existence is certainly present in general, one would have to preserve the variability of covarions over a larger set of substitutions. Hence a larger v is a necessity for what would appear to be a more biologically realistic model" (Model I or II, with u = 0.416, compared to his model).

Comparison with Holmquist, Cantor, and Jukes' REH Interactive Model

The random evolutionary hit (REH) interactive stochastic model also treats evolution as a stochastic process operating within evolutionary constraints. In this section I will compare the predictions from the interactive and covarion models, suggest explanations for the differences, and outline a program to compare these predictions more precisely.

The interactive model uses minimal base substitution data from pairs of present protein sequences to estimate two parameters for the evolutionary path linking two organisms. The first parameter, T_2, is the number of potentially variable sites; each is open to a possible substitution throughout the evolutionary path, so there is no loss of variability as in the covarion model. The second, μ_2, is the mean of a Poisson process determining the number of substitutions at each of the T_2 potentially variable sites. Thus, μ_2 is also the average number of nucleotide substitutions per variable codon. The model uses the genetic code to account fully for synonymous and back substitutions. It assumes that mutations occur, and substitutions are fixed, at random, an approximation of which the authors are cognizant.

The interactive model predicts that each variable codon in cytochrome c has fixed an average of three to four mutations (mean value of μ_2 is 3.41 ± 1.39; Moore et al.,

1976). However, this average is substantially smaller for closely related organisms. Table 4 summarizes the values predicted for the 153 pairs of the 57 species used by Moore et al. (1976) for which a direct comparison of protein sequences gives at most 13 minimal base substitutions (this number was chosen arbitrarily, and all such pairs were included, not just those which would provide nearly independent estimates). The results suggest that the average number of substitutions per variable codon may be 1.5 to 2.0 for such closely related species. This estimate is also supported by the near diagonal entries in Table 5 of Moore et al. (1976).

Table 5 shows that the parameter estimates from this model are quite sensitive to small changes in the data for pairs of closely related organisms. Since the estimates from the covarion model should also be regarded as qualitative, we certainly can expect to make only qualitative comparisons between the model predictions. We will carry out this comparison using the average number of substitutions per variable codon, estimated as the parameter μ_2 in the interactive model.

A probability computation can be done to obtain the value predicted by the covarion model, provided that we make the stronger Assumption 2a that no codon can regain variability. Let S_m be the total number of substitutions which leave fixed or increase the MMD, given an observed MMD=m. Using the methods outlined in Appendix B, which depend on the assumptions used for the covarion model plus the stronger Assumption 2a, the average values of S_m/m were computed for m=1 to m=5 for both

Table 4. Estimates from the interactive model of the expected number of substitutions per variable codon in cytochrome c for closely related species

μ_2	Number of pairs: animals	Number of pairs: plants
0.57-0.88	17	15
1.02-1.51	18	23
1.71-1.98	16	15
2.37-2.54	14	13
2.98	8	7
3.65	4	1
4.14-4.85	3	5
5.12-5.37	3	1
total pairs	73	80

This table records the number of pairs of species within given ranges of values of the parameter μ_2, the expected number of substitutions per variable codon, estimated from the interactive model, as computed by Holmquist (personal communication). Of the 57 species used, only pairs for which a direct comparison of protein sequences gives at most 13 minimal base substitutions were tabulated. All such pairs were included for which the ratio of codon pairs with MMD=2 to MMD=1 is greater than zero and less than one (otherwise, a different and somewhat arbitrary method is used to compute μ_2), not just those which would provide nearly independent estimates. The "animal" species are mostly mammals, but include a few fish, birds, reptiles, and flies. (The values of μ_2 given are in fact too large, as they were based on the methods given in Jukes and Holmquist (1972); see the explanation in Table 5. For example, the class 1.71 to 1.98 should be 1.50 to 1.70; 2.98 should be 1.98 - 2.37; 3.65 should be 2.98)

Table 5. Parameter estimates from the interactive model for given minimum mutation distance data. (Holmquist, personal communication)

Total MMD	Number of pairs of codons with given MMD			μ_2	T_2
	0	1	2		
4	101	2	1	1.98	4.2
5	100	3	1	1.50	6.4
6	99	4	1	1.21	9.0
8	97	6	1	0.88	15.3
	98	4	2	2.37	7.9
10	95	8	1	0.69	23.4
	96	6	2	1.70	12.0
11	94	9	1	0.63	27.8
	95	7	2	1.50	14.4
	96	5	3	2.98	9.8

The definitions of μ_2 and T_2 are given in the text. Only data for which the ratio of the number of codon pairs with MMD=2 to the number with MMD=1 is between zero and one are given, since only for such data can accurate parameter estimates be made. These values of μ_2 and T_2 were calculated from equations (13), (15), and (3) and Table 1 in Holmquist (1978) to avoid the statistical bias in the original paper of Jukes and Holmquist (1972), which described the calculation of these quantities

Table 6a. Mean values of the total number of substitutions per MDI substitution for selected parameter values

	Model I, u = 0.416				Model I, u = 0.300	
	c = 3		c = 4		c = 3	
m	v = 0.40	v = 0.70	v = 0.40	v = 0.70	v = 0.40	v = 0.70
2	1.67	1.73	1.60	1.65	1.80	1.90
3	1.69	1.79	1.61	1.68	1.84	2.03
4	1.69	1.78	1.61	1.71	1.84	2.02
5	1.68	1.75	1.61	1.72	1.84	1.97

The entries in the table are the mean values of S_m/m, where S_m is the total number of nucleotide substitutions which do not decrease the MMD during m MDI substitutions, subject to the assumptions given in the text. For both Models I and II and all values of c, v, and u, the mean value of S_1 is 1.51. $E(S_m)$, the mean value of S_m, was computed using the methods given in Appendix B. $E(S_m)$ increases as v and u increase, and decreases as c increases, and it eventually decreases as m increases. For example, for Model I with u = 0.416, c = 3, and v = 0.70, $E(S_{10})/10 = 1.55$, and $E(S_{15})/15 = 1.41$. The results for Model II are similar, except that the values of $E(S_m)/m$ are smaller (by at most 0.16 and 0.30 for u = 0.416 and u = 0.300, respectively)

models I and II, using c=3 and c=4, and values of v between 0.40 and 0.70. A sample of the results is given in Table 6a. For u=.416, the means were between 1.5 and 1.7; for u=0.30, which suggests what the model would predict if back substitutions were allowed

Table 6b. Mean number of total nucleotide substitutions per variable codon predicted by the covarion model

c	v	mean	SE	S/VC	mean	SE	S/VC
		Model I, u = 0.416			Model I, u = 0.300		
3	0.40	1.65	0.074	1.37	1.77	0.139	1.47
	0.70	1.71	0.011	2.86	1.89	0.211	3.15
4	0.40	1.59	0.004	0.88	1.67	0.009	0.93
	0.70	1.65	0.008	1.84	1.78	0.158	1.98
		Model II, u = 0.416			Model II, u = 0.300		
3	0.40	1.50	0.001	0.83	1.54	.013	0.85
	0.70	1.62	0.006	1.81	1.73	.129	1.93
4	0.40	1.48	0.002	0.62	1.51	.009	0.63
	0.70	1.57	0.004	1.31	1.65	.080	1.37

The mean given is the mean of the values of $E(S_m)/m$ for $1 \leqslant m \leqslant 5$; SE is the sample standard error of this mean. S/VC is an approximation to the total number of substitutions per variable codon, computed as the mean divided by c(1-v) and (c-1)(1-v) for Models II and I, respectively

(see the preceding section), the means were between 1.5 and 1.9. A rough estimate of the average number of nucleotide substitutions per variable codon is the mean of S_m/m divided by (c-1)(1-v) or c(1-v), the average number of new variable codons after each substitution for Model I or II, respectively; this is denoted by S/VC in Table 6b.

The data in Table 6b show that, with this stronger assumption against regaining variability, the covarion model predicts about 0.9 to 2.0 total substitutions (including those leaving the MMD unchanged) per variable codon. This prediction of μ_2 is less than that from the interactive stochastic model, but by no more than a factor of two. However, under Assumption 2, the weaker restriction on regaining variability, the estimate of μ_2 would increase, since there would be fewer *new* covarions which had never been variable. If the number of MDI substitutions allowed is not very large, say at most 10 to 15, and a substantial proportion of all codons are potentially variable, a covarion set of about four codons and turnover rate of about 50% imply that few codons would regain variability by chance alone, so the estimate of μ_2 from the covarion model would not increase very much.

To show that this comparison is exactly what we should expect if both models are valid (this would mean that, although the underlying assumptions are different, these differences are not too significant), first note that one would expect the interactive stochastic model to predict a larger set of potentially variable codons at a given time than the covarion model. In the stochastic model, the variable codons are those able to fix a substitution in *either* leg of the phylogenetic tree. Thus, if cytochrome c evolved independently in two paths from a common ancestor, during a short time interval we would expect the interactive model to predict T_2 to be about twice the predicted covarion size.

However, during the evolution of two fairly closely related species (MMD at most about 15) we would expect substantial parallelism[3], and, as a result, it is easy to see that the covarion model predicts a (possibly substantially) larger set of potentially variable codons, and hence a smaller number of substitutions per variable codon. For example, suppose the phylogenetic tree indicates that two species with MMD=10 have diverged from a common ancestor, with MMD's on the two legs of m=4 and m=6. Suppose that there have been T total substitutions on the two legs (corresponding to the product $\mu_2 T_2$ in the interactive model), distributed on the legs proportionally to the MMD of each leg. For Model I, under the stronger assumption on regaining variability the average number of potentially variable codons during m MDI substitutions is m(c-1)(1-v), hence on the leg with MMD=4 the number of substitutions per variable codon (S/VC) is $T \cdot (4/10) \div (m(c-1)(1-v)) = T/(10(c-1)(1-v)$. On the leg with m=6, S/VC is the same. Averaging these equal predictions for m=4 and m=6 shows that the covarion model estimate of μ_2 will be T, the total number of substitutions, divided by the average number of potentially variable codons during a number of MDI substitutions equal to the total MMD of the two species. Since the covarion model does not account for parallelism, we should expect the estimated number of potentially variable codons in the covarion model to be at least that (T_2) from the interactive model, but less than twice as great.

The interactive model will predict a larger number of total substitutions, since it takes back substitutions into account. However, back substitutions (those decreasing the MMD) would occur relatively rarely by chance alone, as discussed in the derivation of the models, and seem unlikely to occur often through selection in closely related species. Similarly, although codons might regain variability relatively infrequently during relatively short evolutionary time periods, this possibility may increase the estimated value of μ_2.

Therefore, the estimate of μ_2 from the covarion model should be less than that from the interactive model, but by no more than a factor of two. This is exactly what was found. Thus, both models may be valid.

There are two additional points worth noting concerning comparing these models. First, it seems likely that the comparisons should be made in terms of the estimates of μ_2. The estimate of T_2 depends strongly on the MMD in both models, and it would probably be very difficult to estimate the covarion parameters from μ_2 and T_2. Second, one would certainly like to have more a precise estimate of the parameter v in the covarion model than it has been possible to give here, and perhaps a better model, to obtain a more valid comparison.

There are several obvious ways to improve the covarion model. One could use as data to be fit the number of residues with minimal one- and two-base changes in the maximum parsimony tree computed using Moore's augmented distance method. The model itself could be improved by treating codons with MMD=2 accurately, thus eliminating Assumption 5. A more extensive sensitivity analysis should be done, including

---

[3] In their phylogenetic tree for cytochrome c, Fitch and Margoliash (1968) found 41 parallel substitutions in a total of 230. In their reconstruction of the ancestral myoglobin sequence using a minimum mutation distance cladogram, Romero–Herrera (1973) found 74 parallel substitutions at 24 positions in a total of 147 substitutions.

the effect of increased augmented distances on the estimates, for the distances may increase as more proteins are sequenced. The minimum mutation distance of pairs of species in the maximum parsimony tree should be restricted instead of generalizing the covarion model to allow codons to regain variability. The latter alternative would make the model more complicated and could involve the introduction of an additional parameter, making the data easier to fit but increasing the difficulty in minimizing the quantity analogous to WS (c,v).

The test of the consistency of these two models should be conducted by comparing estimates of the parameter μ_2 for similar classes of pairs of protein sequences (e.g. mammals, mammals-birds, plants), since the evolution of cytochrome c is not uniform in different groups of organisms (Moore et al., 1976), and other proteins. If both models are accepted as valid, it might be possible to use the respective model estimates of μ_2 to estimate how often codons regain variability for "small" minimum mutation distances. Finally, it would be highly desirable to be able to estimate the covarion parameters from those in the interactive model, since the latter are easier to obtain.

At present, the interactive model is probably to be preferred to the covarion model. It uses known (not reconstructed) data, accounts fully for back substitutions, allows the user to obtain parameter estimates easily and cheaply, provides a direct estimate of the rate of evolution, and can fit the observed data for distantly related proteins well (Jukes and Holmquist, 1972; Moore et al., 1976; Holmquist et al., 1976).

Appendix B

Computation of S_m

Let S_m be the total number of nucleotide substitutions on a leg of the phylogenetic tree containing m MDI substitutions, excluding only back substitutions; covarions with MMD=2 will be treated as if their MMD were one, as in Appendix A. We wish to compute $E(S_m)$, the mean value of S_m. Let T_i be the time at which the *ith* MDI substitution occurs, with the start of the leg at $T_0=0$; $T_m + t_m$, the time corresponding to the end of the leg; N_i, the number of nucleotide substitutions in the time interval $(T_i,T_{i+1}]$; and e_m, the number of substitutions in the time interval $(T_m, T_m + t_m]$. Thus $S_m = N_0 + N_1 + \ldots + N_{m-1} + E_m$, so

$$E(S_m) = \sum_{i=0}^{m-1} E(N_i) + E(e_m). \qquad (9)$$

First we deal with the final term in (9). It is standard to model the random quantities e_m and N_j by Poisson processes (Parzen, 1962, p. 29). Clearly $0 \leqslant t_m < T_{m+1}-T_m$. Assume that t_m is uniformly distributed on the interval $[0,T_{m+1}-T_m)$. Then $E(t_m)$ is $(T_{m+1}-T_m)/2$ (see e.g. Parzen, 1962, p. 14). Since the m+1 *th* MDI substitution cannot occur during the time interval $[T_m,T_{m+1})$, in which t_m lies, the expected number of substitutions during this interval is $E(N_{m-1})-1$. Furthermore, since we have Poisson processes, $E(e_m)$ and $E(N_{m-1})$ are proportional to $E(t_m)$ and $T_{m+1}-T_m$, respectively, with the same proportionality constant, so

$$E(e_m) = (E(N_{m-1}) - 1)/2. \tag{10}$$

Now we compute $E(N_i)$. First, recall that, if r_{ik} is the probability that N_i equals k, then

$$E(N_i) = \sum_{k=0}^{\infty} k\, r_{ik}.$$

Now let a_i be the probability that a nucleotide substitution occurring during time $(T_i, T_{i+1}]$ does not increase the MMD; then

$$a_i = \Pr(\text{synonymous substitution in codon with MMD=0}) + \Pr(\text{hit on codon with MMD=1 doesn't increase the MMD}). \tag{11}$$

Since r_{ik} is the probability that it takes exactly k experiments to obtain the first "success" in a sequence of Bernoulli trials, $r_{ik} = a_i^{k-1}(1-a_i)$ and it follows that

$$E(N_i) = (1 - a_i)^{-1}. \tag{12}$$

We can use a Markov chain (see Appendix A) to compute the probabilities a_i using Equation (11). From Table 8 of Holmquist et al. (1972), the probability of a synonymous substitution is 0.255 (excluding chain-terminating codons). Therefore, given that n_i=k (there are exactly k covarions with MMD=0 after the *ith* model step), Equation (11) implies that $a_i = .255\,(k/c) + (1-u)(1-k/c)$, where u is defined in the specification of the model. Since we do not know k, we use conditional probability: Let

$$\pi_{ik} = \Pr(n_i=k)$$

Then

$$a_i = \sum_{k=0}^{c} \left\{ .255k/c + (1-u)\frac{c-k}{c} \right\} \pi_{ik} \tag{13}$$

The π_{ik} can be computed as a Markov chain; we use the notation of Appendix A, but now

$$p_{jk} = \Pr(n_{i+1} = k | n_i = j). \tag{14}$$

Let E_i be the event that the *ith* MDI substitution is at a covarion with MMD=0, and F_i, that it is at a covarion with MMD=1 or 2. Note that

$$p_{jk} = \Pr(n_{i+1} = k \text{ and } E_i | n_i=j)\cdot p_j + \Pr(n_{i+1} = k \text{ and } F_i | n_i=j)\cdot(1-p_j) \tag{15}$$

where p_j is the probability defined in Equation (1). The first conditional probability in (15) is exactly (7) or (8) in Appendix A, for Model I or II, respectively. Using exactly the same reasoning used to derive Equations (7) and (8), the second conditional probability in (15) is easily shown to be, for $k \geqslant j-1$

$$\binom{c-j-1}{k-j} v^{c-k-1} (1-v)^{k-j} \quad \text{for Model I}$$

$\binom{c-j}{k-j} v^{c-k}(1-v)^{k-j}$ for Model II

and zero for $k \leqslant j - 2$, or if k=c in Model I.

Finally, to compute the π_{ik} in (13), let $\pi^{(i)}$ be the row vector of length c+1 of the π_{ik}. Note that $\pi^{(0)} = (0,...,0,1)$ and that, for Model I, $\pi_{ic} = 0$ if $i \geqslant 1$. Let P be the matrix of transition probabilities p_{jk} in (14), computed using (15). Then, just as in Appendix A, $\pi^{(i)} = \pi^{(0)}P^i$.

Thus, once the π_{ik} have been computed, the a_i can be computed from (13); the $E(N_j)$, from (12), and $E(e_m)$, from (10); and finally, $E(S_m)$, from (9).

Acknowledgements. I thank Richard Holmquist and Walter Fitch for their criticism of a preliminary version of this paper; Jerry Fields, Tel Aviv University, for discussions on the chi-squared and parameter estimation; Thomas Kinraide, Colorado College, for discussion on protein function; and Ms. Beverly Price, for her skillful typing of the manuscript. I am most grateful to The Research Corporation for support under a Cottrell College Science Grant. Part of this work was done while the author was visiting the Department of Statistics, Tel Aviv University, which provided computer time.

References

Fitch, W.M. (1971). J. Mol. Evol. **1**, 84-96

Fitch, W.M. (1976). J. Mol. Evol. **8**, 28

Fitch, W.M., Margoliash, E. (1968). Brookhaven Symp. in Biol. **21**, 217-242

Fitch, W.M., Markowitz, E. (1970). Biochem. Genet. **4**, 579-593

Holmquist, R. (1976). In: Molecular Anthropology, (M. Goodman, R.E. Tashian, J.H. Tashian, eds.) New York: Plenum Publishing Co

Holmquist, R. (1978). J. Mol. Evol. **11**, 361-374

Holmquist, R., Cantor, C., Jukes, T.H. (1972). J. Mol. Biol. **64**, 145-161

Holmquist, R., Jukes, T.H., Moise, H., Goodman, M., Moore, G.W. (1976). J. Mol. Biol. **105**, 39-74

Jukes, T.H., Holmquist, R. (1972). J. Mol. Biol. **64**, 163-179

Lancaster, H.O. (1969). The chi-squared distribution. New York: Wiley

Lin, D.K., Niece, R., Fitch, W.M. (1973). Nature **241**, 533

Moore, G., Goodman, M., Callahan, C., Holmquist, R., Moise, H. (1976). J. Mol. Biol. **105**, 15-37

Parzen, E. (1962). Stochastic processes. San Francisco: Holden-Day

Pettigrew, G.W. (1973). Nature **241**, 531

Romero-Herrera, A.E. (1973). Nature **246**, 389-395

Received July 17, 1978