J Mol Evol (2001) 53:711–723
DOI: 10.1007/s002390010258

JOURNAL OF MOLECULAR EVOLUTION

Mathematical Elegance with Biochemical Realism: The Covarion Model of Molecular Evolution

David Penny,[1] Bennet J. McComish,[1] Michael A. Charleston,[2,*] Michael D. Hendy[2]

[1]Institute of Molecular BioSciences, P.O. Box 11222, Massey University, Palmerston North, New Zealand
[2]Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand

Received: 13 February 2001 / Accepted: 22 May 2001

Abstract. There is an apparent paradox in our understanding of molecular evolution. Current *biochemically* based models predict that evolutionary trees should *not* be recoverable for divergences beyond a few hundred million years. In practice, however, trees often appear to be recovered from much older times. *Mathematical* models, such as those assuming that sites evolve at different rates [including a Γ distribution of rates across sites (RAS)] may in theory allow the recovery of some ancient divergences. However, such models require that each site maintain its characteristic rate over the whole evolutionary period. This assumption, however, contradicts the knowledge that tertiary structures diverge with time, invalidating the rate-constancy assumption of purely mathematical models. We report here that a hidden Markov version of the covarion model can meet both biochemical and statistical requirements for the analysis of sequence data. The model was proposed on biochemical grounds and can be implemented with only two additional parameters. The two hidden parts of this model are the proportion of sites free to vary (covarions) and the rate of interchange between fixed sites and these variable sites. Simulation results are consistent with this approach, providing a better framework for understanding anciently diverged sequences than the standard RAS models. However, a Γ distribution of rates may approximate a covarion model and may possibly be justified on these grounds. The accurate reconstruction of older divergences from sequence data is still a major problem, and molecular evolution still requires mathematical models that also have a sound biochemical basis.

Key words: Covarion model — Hidden Markov model — Rates across sites — Role of models in science

Introduction

The earliest models for the evolution of sequences assumed that all variable sites evolved at the same rate. Two approaches were introduced early in the study of molecular evolution to relax this assumption and to allow for rate heterogeneity between sites. One was the well-known approach of fitting a probability distribution of rates (Uzzell and Corbin 1971); the other was the covarion model (Fitch and Markovitz 1970; Fitch 1971). Under the Uzzell and Corbin [rates-across-sites (RAS)] model, each site has a characteristic (or intrinsic) rate which is maintained over the whole time period being studied. Sites differ in these intrinsic rates and this can be modeled by a probability distribution; for a detailed analysis see Chang (1996). This general approach, especially the Γ distribution (see Yang 1996), has been well developed mathematically, but several other distributions (e.g., Waddell et al. 1997) and empirically measured distributions (Van de Peer et al. 1996) have been used.

These distributions have desirable mathematical properties in that they require only one or two additional

Correspondence to: David Penny; E-mail: d.penny@massey.ac.nz
**Current address:* Zoology Department, University of Oxford, Oxford, England

parameters, irrespective of the length of the sequences. However, a problem with these RAS models is that there is no biochemical explanation as to how each site maintains its own intrinsic rate over long evolutionary periods. This is a problem because the mathematics used for the RAS models does require that each site maintains its intrinsic rate over the entire time period being studied (see Chang 1996; Tuffley and Steel 1997). Thus RAS models require that each site maintains its own intrinsic rate throughout evolution; our knowledge of biochemistry contradicts this (see later). Some features of the structural evolution of proteins are considered next, to illustrate the potential biochemical basis for evolutionary models.

Structural Evolution of Proteins

It is difficult to find a biochemical mechanism that would maintain the same intrinsic rate of evolution at each site, irrespective of whether the gene was in eukaryotes, archaea, or eubacteria—or, indeed, within thermophiles or mesophiles. In fact, biochemical information predicts the opposite; homologous sites in widely different lineages should vary in their rates. The following examples show that one of the strongest conclusions from structural biology is that the three-dimensional (3-D) structure of proteins varies during evolution. This has been demonstrated, for example, by changes in the root-mean-square (rms) difference in the position of the α carbon atoms (C_α) along the backbone of the 3-D structure of a protein (measured in angstroms). In an important study, Chothia and Lesk (1986, 1987) reported on a variety of proteins and showed that the average rms difference in 3-D structure increased with the sequence divergence—even if considering only the core of the proteins. The effect was nonlinear, with increasing difference in 3-D structure at higher sequence divergence. Similar effects were found using the structural alignment score (Levitt and Gerstein 1998). In another wide-ranging study, Pascarella and Argos (1992) also showed changes in structure resulting from 714 insertions/deletions (see, for example, their Fig. 6). Similarly, Carrugo and Argos (1997) have discussed, for a wide range of enzymes, the evolution of the 3-D structure of nucleotide-binding domains. They find examples of both divergence and convergence of the domains.

There have also been many studies on specific proteins. For example, when comparing the X-ray crystallographic structures of a fish and human hemoglobin the rms difference is 1.4 Å, though the closeness of the match varies throughout the protein (Camardella et al. 1992; e.g., their Fig. 8). Another example is studies showing changes in 3-D structure between repeated units of a protein. The subunits had diverged in structure over time even if the units were identical initially in their 3-D structure. An example is the "regulator of chromosome condensation," namely, RCC1. It has a seven-blade propeller structure, but the seven repeating units deviate slightly in 3-D structure (Renault et al. 1998, Fig. 3).

To continue with other types of studies, in an artificial evolution experiment, Spiller et al. (1999) find variants of an esterase that are optimal under different environmental conditions. They show that these variants also have differences in 3-D structure and describe their results in terms of a fitness landscape of 3-D structure through which the enzyme evolves. Finally, a recent study (Fisher et al. 2000) uses site-directed mutagenesis and NMR to test (and confirm) predictions from the covarion model about the effect of specific mutations on structure. A recent overview of protein structural evolution is available (Lesk 2000, Chaps. 5 and 6).

The overriding conclusion is that, although a few essential sites may be invariable over long periods of evolutionary time, most sites do change their functional environment during evolution. As such, the functional constraints on sites are expected to change. This is perhaps one of the best-substantiated facts of structural biology—individual amino acid sites are not in the same environment over all of evolution. Thus, although the probability distribution (RAS) approach has desirable mathematical properties, the requirement for each site to maintain its characteristic rate throughout evolution (Chang 1996; Tuffley and Steel 1997) is not in agreement with our current understanding of biochemical processes. It might have been hoped that RAS models could handle the biochemical situation if the same distribution of rates were maintained throughout evolution—even though individual sites varied in their rate class. For example, it could be that the proportion of fixed sites stayed constant, even if individual sites were variable or fixed in different parts of the tree. Similarly, another proportion of sites may be evolving at no more than 10% of the maximum rate, but again the actual sites varied their rate—though the overall distribution was still maintained. Unfortunately, the mathematical proofs (see earlier) for the RAS model require that each site maintain its intrinsic rate over the whole tree. Thus the biochemical results contradict (disprove?) an essential assumption of the RAS approach.

Covarion Model

The covarion substitution model (Fitch and Markovitz 1970; Fitch 1971) was introduced at the same time as the RAS model of Uzzell and Corbin (1971). The covarion model posits that, although some sites in a macromolecule are critical to function and can never change, most sites switch between being free to evolve in some taxa and being fixed in others. This switch during evolution would result from slight changes in secondary and ter-

tiary structures referred to above. For example, in an early study of cytochrome *c* the overall rate of evolution was about 10% of the neutral rate, consistent with 10% of sites being free to vary at any one time. However, in mammals, 15% of sites had changed, but over a wide range of eukaryotes, about 70% of positions had changed (Fitch and Markowitz 1970). The conclusion drawn was that

> because of the structural restraints imposed by functional requirements, mutations that will not be selected against are available only for a very limited number of positions. . . . However, as such acceptable mutations are fixed they alter the positions in which other acceptable mutations may be fixed. Thus, only about ten codons, on the average, in any cytochrome *c* may have acceptable mutations available to them but the particular codons will vary from one species to another. We shall term those codons at any one instant in time and in any given gene for which an acceptable mutation is available as the *concomitantly variable codons.* (Fitch and Markowitz 1970, p. 585)

"Covarion" is a contraction of *co*ncomitably *vari*able cod*on*s, and of course, the principle can be applied at the nucleotide level (Fitch 1986), as well as for proteins. We prefer *not* to use the term covariotide for the nucleotide version; the underlying concepts are identical irrespective of the number of character states, and it is highly undesirable to increase new terms for every possible variant (Penny 1993).

The covarion model is an extension of the earliest explanations for the differing rates of evolution between proteins (e.g., Dickerson 1971). (Here we are considering differences in the rates of evolution between proteins, not differences between sites within a protein.) We call these earliest explanations of differential rates in protein evolution the Kimura–Dickerson model. Under this biochemical model, changes in sequences are both stochastic and neutral (Kimura), and variation in the proportion of unconstrained ("free to vary") sites accounts for the different rates between proteins (Dickerson). Several authors had suggested that differences in the number of sites free to vary within a protein accounted for the differences in rates between proteins. Dickerson (1971), an early structural biologist also interested in evolution, expresses the idea clearly in relation to 3-D structure.

Figure 1 illustrates this simple biochemical model. Each site free to change evolves at the same rate—for nucleotides this is independent of whether it is the first, second, or third position in the codon. Under this model the first two positions have more sites constrained, consequently their *average* rate is lower (though their variance would be higher). In agreement with this, cases

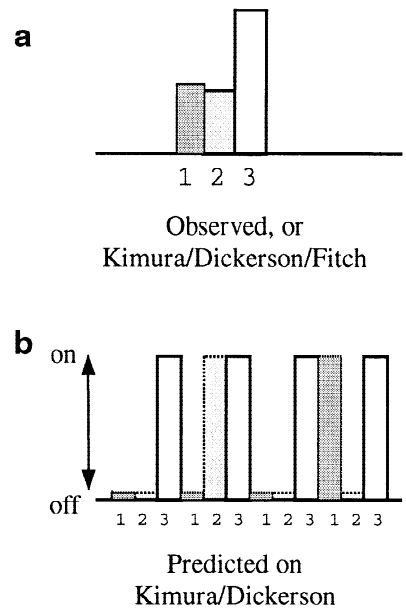

Fig. 1. The problem of explaining different evolutionary rates within the codon. As shown in **a**, the average rate of change usually varies between codon positions, with the second position the slowest and the third the fastest. However, simple biochemical models assume that most nucleotide changes are either neutral or lethal, and therefore sites are either variable or constant ("on" or "off" in **b**). Thus, in these simple models the first and second sites evolve more slowly on *average*, but each variable site saturates at the same rate (see Griffiths 1997). This predicts that the first and second positions of a codon are basically no more reliable than the third codon positions, although the experience of most researchers is that first and second sites are more reliable for older divergences.

have been reported where there is little difference in the rate of saturation at the three positions in the codon, for example, in cytochrome *b* (Griffiths 1997). This observation is expected only when there are no changes in 3-D structure (and, consequently, in structural constraints) of a protein (remembering that under this model the slower *average* rate at the first and second positions is because fewer sites are variable). That equal rates of saturation at the three codon positions is found at all is strong reinforcement that the basic biochemical model (the Kimura–Dickerson model) is part of the evolutionary process, even if incomplete by itself. Under more complex (and realistic) models, this equal rate of saturation at all three codon positions need not occur.

The covarion model is an extension of this basic Kimura–Dickerson model where some sites switch between "on" and "off" in different parts of the tree. Thus it may be called a Kimura–Dickerson–Fitch model. This name is not intended to replace the well-established covarion name; rather it is to show that the model is made up of components, thus making it easier to analyze. Karon (1979), using Fitch's cytochrome *c* data, improved the original model to account more fully for the redundancy in the genetic code and used more robust statistical methods to fit the model to the data. More recently, Fitch and Ayala (1994) reported that the rate of evolution of

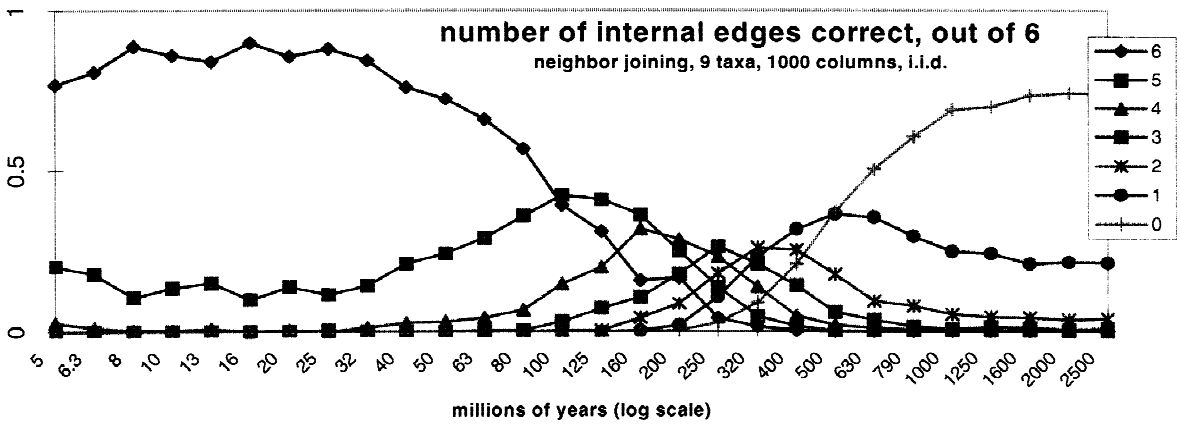

Fig. 2. Expected accuracy versus time of divergence. Data were generated under a Kimura 2-ST model for nine taxa for 5–2500 million years (*x*-axis). Neighbor-joining (Phylip) was used to infer a tree from each data set. The proportion of correct internal edges (branches) is shown on the *y*-axis, ranging from 6 internal edges (branches), the fully correct tree, to 0, where all internal edges were wrong. Each point represents 1000 simulations with sequences 1000 nucleotides long. For shorter times the correct tree (or a tree with only one error) was found in virtually all trials; after the equivalent of about 400 Mya the fully correct tree was not recovered under this model of evolution.

superoxide dismutase (SOD) was consistent with a molecular clock if modeled by a covarion process. Although not commented on, their model was similar to a hidden Markov process (Elliot et al. 1995); see below. Later work (Miyamoto and Fitch 1995) suggested that the covarion model gave a better fit to the data than arbitrary mathematical models, such as the commonly used Γ distribution of RAS. Similarly, a new quantitative test (Lockhart et al. 1998) shows that for their data a covarion model fits the data better than a RAS model—or more accurately, their test rejects any RAS model where each site always has the same rate. Several other authors have limited discussions of the covarion model (e.g., Koonin and Gorbalenya 1989; Marshall et al. 1994). A review of over a hundred papers that mention covarions revealed only one author (Gillespie 1988) who appeared to disagree with the covarion hypothesis.

Despite its sound biochemical basis and its potential importance for evolutionary studies, the covarion model has not been fully developed; from a statistical viewpoint it has far too many parameters to be useful. If most amino acid positions are constant over some portions of the tree and variable in others, then it appears that one could include as many parameters as desired "in order to fit the data to the model"! In general, invoking more and more parameters weakens the power of any model (see Steel and Penny 2000). Indeed in the case of evolutionary trees, Steel et al. (1994) proved that with enough variability of rates between sites, any data could, in principle, be derived from any tree. Thus the covarion model had the converse properties to the probability distribution (RAS) approach; the covarion model had a reasonable biochemical foundation but (as originally formulated) lacked the required mathematical properties. However, before proceeding with an implementation of the covarion model, we study the problem of saturation in the basic Kimura–Dickerson model.

Saturation in the Basic Biochemical Model

It is expected that the basic biochemical (Kimura–Dickerson) model, which assumes that a site is always in the same rate class, would lead to sequences saturating relatively quickly during evolution. This is most easily illustrated by simulation. From previous computer simulations for a variety of methods, recovering evolutionary trees is inaccurate when there has been, on average, more than about one change per site (Charleston et al. 1994). However, most simulation studies use neither defined time periods nor known rates of molecular evolution. For the present study, we use the neutral rate of evolution for unconstrained sites as 0.5% per site per million years (Li 1997, p. 75; Page and Holmes 1997, p. 239). There is then an average of one change per site every 200 million years. Using a measured rate allows the essential linking of simulation results to elapsed time.

To illustrate this problem of saturation for ancient divergences with simple biochemical models (that is, neither covarion nor RAS models), we ran 300,000 simulations on randomly selected nine-taxon trees. The time periods for the simulation increased logarithmically so that the expected numbers of changes per site (which is linear with time) increased from 0.025 to 12.5. This is equivalent to 5 to 2500 million years of evolution and thus gives a relationship between simulation results and evolutionary time. Individual lengths of edges (branches) on the tree were selected randomly, with internal edges being no more than one-half of the longest external edge. Data sets were generated for two-state characters using the Hadamard conjugation (Hendy and Charleston 1993). Data were generated for 100–1000 variable sites (that is, excluding any invariable sites). Trees were inferred from each data sample by a variety of methods.

The results in Fig. 2 show the expected error in recovering the correct tree for nine taxa (six internal edges)

and are for 1000 variable sites using neighbor-joining (Felsenstein 1997)—which is consistent under this model. The increasing time of divergence (*x*-axis) is equivalent to 5 to 2500 million years of evolution at this mutation rate. The *y*-axis is the proportion of internal edges (branches) inferred correctly. For shorter times it is expected that either the entire tree (six internal edges) will be recovered correctly or a tree with no more than a single error. However, as the overall time increases, the expected number of correct branches decreases (relatively rapidly) until eventually the inferred tree is expected to have all six internal edges wrong! Note that even at the extreme the results are still better than random; the method is consistent and would eventually get the correct tree if extremely long sequences were available.

Our conclusion is that under a simple biochemical model (the Kimura–Dickerson model), together with realistic rates of evolution, it should be difficult or impossible to recover trees accurately after divergences of more than about 300–400 million years. This conclusion comes from the present work (where only neighbor-joining results are shown), those of Charleston et al. (1994), the covariance matrix calculated by Hadamard transforms (Waddell et al. 1994), and the length of sequence that can be required even for maximum likelihood (Steel and Penny 2000). Although in practice the long edges on the tree may be broken up by other taxa, the present example illustrates the problem of current models for ancient divergences. Under these standard models, there is no justification for expecting correct results for ancient divergences.

In contrast with theory, many researchers (because of agreement with other data) appear confident with many aspects of evolutionary trees for older divergences. This is the basic problem—simple biochemical models with neutral changes, and sites always having the same rate of change, predict that ancient divergences would be poorly handled by current methods. This difference between theory and practice must be addressed.

In defense of theory, it is well known that the order of divergences of the main avian and mammalian groups is still controversial. This result is consistent with theory and is found even for eutherian mammals that diverged within the last 130 million years (relatively recently on the scale in Fig. 1). There has been good progress in resolving this question (Waddell et al. 1999) but the point at issue is, why are we more confident of much older divergences when we know that we cannot guarantee more recent ones?

In defense of practice, it is repeatedly found that there is considerable agreement between different data sets (both molecular and morphological). Bird sequences do not come out among mammals or invertebrates, mosses among flowering plants or fungi, and so on. There are many difficulties with differences between data sets (especially with the oldest divergences) but there is no suggestion (as in Fig. 2) that all internal edges of a tree are incorrect. Nevertheless, there are major difficulties between data sets for ancient divergences. It is difficult to see why researchers are so confident in their results when the relatively recent divergences within mammals, birds, or flowering plants are only now being resolved. The covarion model offers a possible resolution to this fundamental problem of molecular evolution. But first we consider a range of alternatives as to why some sites evolve more slowly.

Alternative Models

A standard answer is just that some sites "change more slowly" or that "some sites are more constrained." However, this is a *description* of what is observed, not an *explanation.* What mechanisms would result in a site in a protein "changing more slowly"? Possibilities include the following:

(1) mutation rates are lower at first and second positions;
(2) mutations fixed at first and second positions are slightly deleterious and therefore are less likely to be fixed (giving a slower rate);
(3) fewer mutations are viable at first and second positions (but sites are always variable or always fixed—the Kimura–Dickerson model);
(4) first and second positions can switch between being "on" and being "off" with slight changes in secondary and tertiary structure (a covarion model); and
(5) a new model will explain the observation.

A difference in mutation rate between codon positions in a single gene is not considered likely as a general explanation (though mutation rate may vary in different parts of the genome). DNA polymerases, proofreading, and error correction all function independently of the reading frame. The second explanation is the fixation of some slightly deleterious mutations in a population. Such sites will change more slowly than the neutral rate—because negative selection still eliminates some mutations. By itself, this is not a general solution because it implies a slow continual decline in fitness over time. A different version of this model is due to Zuckerkandl (1976), where positive selection was invoked for fixation of virtually all changes. In this model proteins were suboptimal (because some of the desirable properties were incompatible). There was a continual cycle of fixation of new mutations that improved one aspect of the protein but led to decreased functionality in another. Although the model was a selectionist explanation for a molecular clock, it never received direct biochemical support and there was increasing support for the view that most

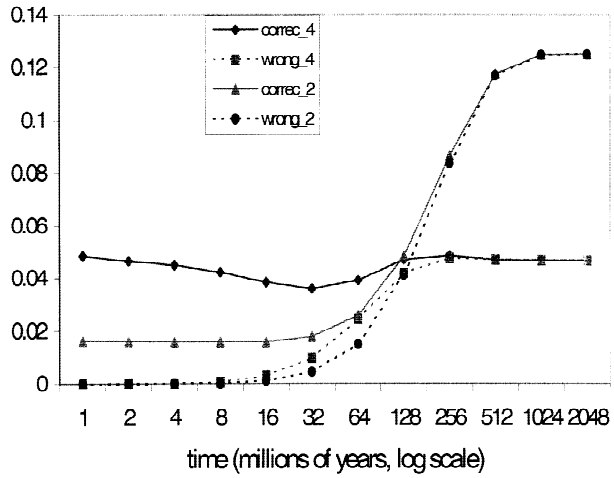

Fig. 3. Rate of saturation when two or four nucleotides are viable at a site, for increasing times. The probabilities of a site supporting the correct tree are correc_2 (for two viable nucleotides) and correc_4 (for four viable nucleotides). Conversely, probabilities for sites supporting the wrong tree are wrong_2 and wrong_4. The results show that sites saturate at about the same rate, irrespective of whether two or four nucleotides are viable. For two viable nucleotides, there are only 8 patterns at a site, compared with 64 when all nucleotides are viable (three of which directly support the correct tree).

changes were neutral. Indeed, positive selection for similar changes on different lineages would only make it harder to recover the correct tree.

We demonstrate here that the third explanation, differences in the number of viable nucleotides at a site, does not work either. The basic Kimura–Dickerson model of molecular evolution allows from one to four nucleotides at a site to be viable. On this explanation, negative selection eliminates from zero to three mutations. If most mutations at a site were lethal, would this maintain a phylogenetic signal longer? This can be checked by calculation.

Consider the possibility that only two nucleotides (or amino acids) were viable at a site. The rate of loss of phylogenetic signal was calculated on a four-taxon tree that was rooted on the central edge and had equal rates of evolution [so there were no problems with inconsistency (Hendy and Penny 1989)]. Using the method of Hendy et al. (1994), calculations were made for all nucleotide changes equally likely (the Jukes–Cantor model) and with 5% change on the internal edge. The external edges (equal to time) were then made longer and longer and are shown as the *x*-axis in Fig. 3. Time is on a logarithmic scale up to 2 billion years (a time routinely used when studying the tree of life). "Correc_2" and "correc_4" represent the probability of a site supporting the correct tree with two, or four, nucleotides viable at that site. Conversely, "wrong_2" and "wrong_4" are the probabilities of a site supporting one of the two incorrect trees (values should be doubled if considering the probability of a site supporting either incorrect tree). Note that with only two of the four nucleotides viable, there are (for $t = 4$ taxa) only 8 possible patterns at a site; there are 64 patterns for four nucleotides. Consequently, as sites become randomized each value converges to 0.125 for two viable nucleotides [1 of 8 (2^{t-1}) patterns] and to 0.0469 for four nucleotides [3 patterns of 64 (4^{t-1})].

In Fig. 3 there is little difference in the rate of randomization (loss of phylogenetic signal) whether two or four nucleotides are viable at a site. Under the parameters used, sites with either two or four nucleotides viable are approaching randomization by 300–400 million years. With only two nucleotides viable there is a 1–3% slower approach to randomization—essentially no difference. Another conclusion from the results in Fig. 3 (not shown) was that the proportion of sites that had not changed, although they were free to vary, decreased to zero relatively quickly. This reinforces the conclusion that sites that are constant for anciently diverged trees are functionally constrained (that is, genuinely invariable). Such sites should not be used when estimating either the number of multiple changes or the nucleotide compositions at other sites that are free to vary. Retaining such sites in an analysis means that the number of changes is underestimated (Palumbi 1989), and even maximum likelihood is no longer consistent (Lockhart et al. 1996; Steel and Penny 2000).

The results in Figs. 2 and 3 illustrate a fundamental problem with inferring ancient evolutionary trees from sequence data; current biochemically based models are not encouraging regarding our ability to recover deep-branching phylogenetic signals. However, our working hypothesis was that, with real sequence data, processes such as the continuous operation of a simple covarion model could make the inference of older divergences more accurate. We now report results that support this expectation. Data were generated under a simple covarion model and then analyzed by standard (noncovarion) models.

Covarion Model with Two Additional Parameters

Given the failure of the first three suggestions for slower average evolution at the first two positions in a codon, we now consider covarion models. These allow sites to vary in their rate of evolution as the 3-D structure of the macromolecule evolves. Such models may be useful if we can combine biochemical realism with mathematical rigor. We have been involved in a long-term study of covarion models and the present work is the background to this wider study. The areas of interest include mathematical analysis (Tuffley and Steel 1997) and a demonstration that sites evolve at different rates over the tree (Lockhart et al. 1996, 1998). The work has led to tests (Lockhart et al. 1998, 2000) that distinguishes a covarion model from one that predicts sites are always in the same rate class. This present paper gives the basic reasoning

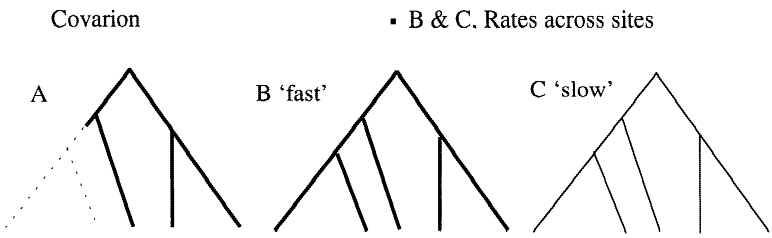

Fig. 4. A difference between the covarion model (**A**) and rates-across-sites models (**B** and **C**). Under the covarion model, a site can change from the standard rate (*solid line*) to zero (*dashed line*), and vice versa. Under the rates-across-sites model (B and C), each site maintains its own intrinsic rate over the whole tree, though there is a distribution of sites evolving at different rates (faster in B, slower in C).

behind our studies and reports results from the application of a partially hidden Markov process to model the covarion process. The model requires only *two* parameters additional to the commonly used Markov models (Tuffley and Steel 1997). It thus solves the main problem in the past, that the original covarion model appeared to require several parameters per site. We give the formal model here for nucleotide evolution but it is readily extended to amino acids.

The hidden Markov model has two main processes. The first is a standard Markov model for molecular evolution and is implemented here with the Kimura (1981) 3ST model—which contains the 2ST and Jukes–Cantor models as special cases. The second (hidden) process has the two additional parameters: φ, the proportion of sites that are free to vary (the covarions); and δ, the rate of interchange between the covarions and sites that are invariable (cannot change because of biochemical constraints). In this simple version, all sites have the same probability of being in either rate class, and thus the model is still stationary and i.i.d (independent and identically distributed). The model is "stationary" in the sense that the basic process and frequencies of rate classes are unchanged over the whole tree. At a variable site, a mutation may be fixed in the population either by random genetic drift (neutral) or by positive selection [although the model has no enhanced rate of fixation for positive selection (see also Ohta and Kimura 1971)]. Whether a site in a sequence is fixed or variable at a particular point in time is unknown—hence the name "hidden" Markov model. The variability status is represented by a superscript: + when the states are free to change (A^+, G^+, C^+, and T^+) and − when they are fixed (A^-, G^-, C^-, and T^-). For proteins, the 40 character states are A^+, C^+, D^+, E^+, F^+, . . . Y^+ for the potentially variable sites and A^-, C^-, D^-, E^-, F^-, . . . Y^- for sites that are invariant at a particular point in time. The rate of interchange between the fixed and the variable states is set to maintain the proportion φ of variable (covarion) and fixed sites (see below).

Figure 4 illustrates the difference between the covarion model (Fig. 4A) and the distribution of RAS model. In the covarion model, a site can change in the tree between at least two rate classes. In the distribution of RAS model, each site (or category) has its own rate, which is maintained over the entire tree, for example, either fast as in Fig. 4B or slow as in Fig. 4C.

Kimura's 3ST model is used by both RAS and covarion models and is described by an instantaneous rate matrix, **K**, which has three substitution types, α, β, and γ; hence the name 3ST. The Hadamard matrix **H** diagonalizes **K**:

$$\mathbf{K} = \begin{array}{c} \\ \text{A+} \\ \text{G+} \\ \text{C+} \\ \text{T+} \end{array} \begin{array}{c} \begin{array}{cccc} \text{A+} & \text{G+} & \text{C+} & \text{T+} \end{array} \\ \begin{bmatrix} * & \alpha & \beta & \gamma \\ \alpha & * & \gamma & \beta \\ \beta & \gamma & * & \alpha \\ \gamma & \beta & \alpha & * \end{bmatrix} \end{array}$$

$$\mathbf{H} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}$$

$$\mathbf{\Lambda} = \mathbf{H}^{-1}\mathbf{K}\mathbf{H} = \begin{bmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 \\ 0 & 0 & 0 & \lambda_4 \end{bmatrix}$$

where $* = -(\alpha + \beta + \gamma)$ (so that the rows sum to zero); $\mathbf{H}^{-1} = \frac{1}{4}\mathbf{H}$; and $\lambda_1 = 0$, $\lambda_2 = -2(\alpha+\gamma)$, $\lambda_3 = -2(\beta+\gamma)$, and $\lambda_4 = -2(\alpha+\beta)$ are the eigenvalues for **K** (for $\alpha > \beta \geq \gamma$). This diagonalization enables the ready calculation of the exponent of **K**, $\exp(\mathbf{K}) = \mathbf{I} + \mathbf{K} + (\mathbf{K}^2/2!) + (\mathbf{K}^3/3!) + (\mathbf{K}^4/4!) + \ldots = \mathbf{H}\cdot\exp(\Lambda)H^{-1}$, and $\exp(\Lambda)$ is the diagonal matrix whose entries are $\exp(\lambda_i)$. The transition matrix $\mathbf{M} = \exp(\mathbf{K}t)$ expresses the probabilities of the substitutions of each type during an interval of time *t* and the values from **M** are used to predict the amount and types of nucleotide changes during evolution.

A basic rate matrix (**M**) for our hidden Markov model is

$$\mathbf{K}' = \begin{bmatrix} \mathbf{K} & 0 \\ 0 & 0 \end{bmatrix} + \delta \begin{bmatrix} -\mathbf{I} & \mathbf{I} \\ kI & -\mathbf{I} \end{bmatrix}$$

where $\mathbf{K}'$ is an 8 × 8 matrix and **K** and **I** are 4 × 4 matrices. To make the model more general (allowing different proportions of fixed and variable sites), we choose a value for *k* so that the proportions of variable and constants sites is constant during evolution. The instantaneous rate matrix ($\mathbf{K}'$) for the hidden Markov model we actually use, based on a Kimura three-parameter model (**K**), is shown in Fig. 5. From this point our use of the rate matrix $\mathbf{K}'$ is standard; for specific values of α, β, γ, δ, and *k*, we take the exponential to get

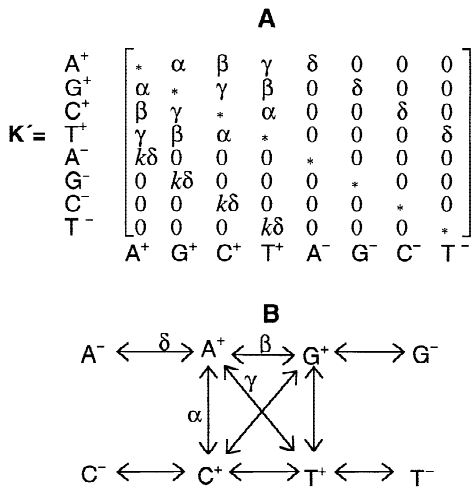

Fig. 5. Parameters for a hidden Markov model for nucleotide evolution. **A** The instantaneous rate matrix **K′**. **B** A graphical representation. The diagonals (labeled *) are given values so that each row of the rate matrix sums to 0. The *arrows* in the graphical representation (B) correspond to the positive entries in the rate matrix. For example, there are four rates to or from A^+ (α, β, γ, and δ) and only one (δ) from A^-.

a Markov transition matrix—which is used for simulating sequence data (see below).

Our working hypothesis was that the covarion model allows older divergences to be recovered more accurately. The trees chosen to test the covarion model (Fig. 6) are inconsistent under uncorrected parsimony (Hendy and Penny 1989) and are especially difficult to recover without accurate corrections for multiple changes. The four-taxon tree $(t_1,(t_2,(t_3,t_4)))$ (Fig. 6A) has one slowly evolving lineage (t_2), and the other lineages fit a molecular clock. The correct tree becomes the longest (not the shortest) tree for uncorrected parsimony. In the five-taxon tree $((t_1,t_2),(t_3,t_4),t_5)$ all lineages fit the molecular clock (Fig. 6B). However, with uncorrected parsimony the correct tree (even with infinitely long sequences) is the 12th longest (of 15). For each tree, the internal part of the tree was fixed, while the external edges were allowed to increase in length, representing longer and longer times.

Results from Modeling the Covarion Process

Our test was the frequency of recovery of the correct tree as the total period of evolution increased. Sequences were generated under a covarion model and then a Kimura model in Phylip (Felsenstein 1997) used to infer trees. The covarion process was based on the Kimura 2ST model with $\alpha = 0.005$ (the transversion rate) and $\beta = \gamma = 0.0025$ (the transition rate) and with equal nucleotide frequencies at the root. For the two additional parameters for the covarion model, φ (the proportion of variable sites) was set at 0.5 and δ (the rate of interchange between fixed and variable sites) was varied. For the results reported here, sequences were of length 1000. DNAML was used for maximum likelihood and NEIGHBOR for neighbor-joining on distances.

The computational time for maximum-likelihood calculations with a thousand simulations for even a single data point on the five-taxon tree is large. Consequently the simulation and tree-building steps were performed on 50 Pentium PC computers running in parallel and controlled remotely over the network. This process was largely automated by using batch and input files to control each set of simulations and to transfer output to the tree-building programs.

Results for the five-taxon tree (Fig. 6B) are shown in Fig. 7 (the four-taxon tree in Fig. 6A has similar results; data not shown). The *x*-axis is time (on a logarithmic scale) and the *y*-axis the probability of correctly recovering the generating tree. Figure 7A is for neighbor-joining and Fig. 7B for maximum likelihood, both using Phylip. The successive curves are for increasing values of δ (the rate of interconversion between fixed and variable sites). The general conclusion from Fig. 7 is that increasing δ increases the ability to recover the tree correctly. $\delta = 0$ is equivalent to 50% of sites being fixed and 50% being variable (φ is 0.5). As δ increases to ∞, the model becomes equivalent to all sites being variable, but at half the rate of change as with $\delta = 0$. If the covarion model is a realistic biochemical description of molecular evolution, then current methods for inferring trees (that is, methods not using a covarion model) are expected to perform better than the results in Fig. 2.

The extent of the improved performance is shown by the observation that the model tree can still be identified after longer periods of evolution—if a covarion process is operating. Increasing the rate of interchange (δ) between fixed and variable sites increases the chance of selecting the correct tree. DNAML was more successful than neighbor-joining with the four-taxon tree, but with the five-taxon tree, the results were more ambiguous. Both, however, did better as δ increased in value. A covarion model with these parameters increases by 50–100% the time over which current methods of tree building are reliable. This need not be the limit for increased performance. Many other combinations of parameters could be tested, though it is preferable to explore theoretical properties first to test predictions more constructively. Note, again, that the covarion process was used only to generate the data, not when inferring the tree from the data.

Consequences for Ancient Divergences

It is clear that simulation studies need to use both realistic times and real rates of evolution (not rates averaged across variable and invariable sites). Most simulation studies use only relatively short evolutionary periods and

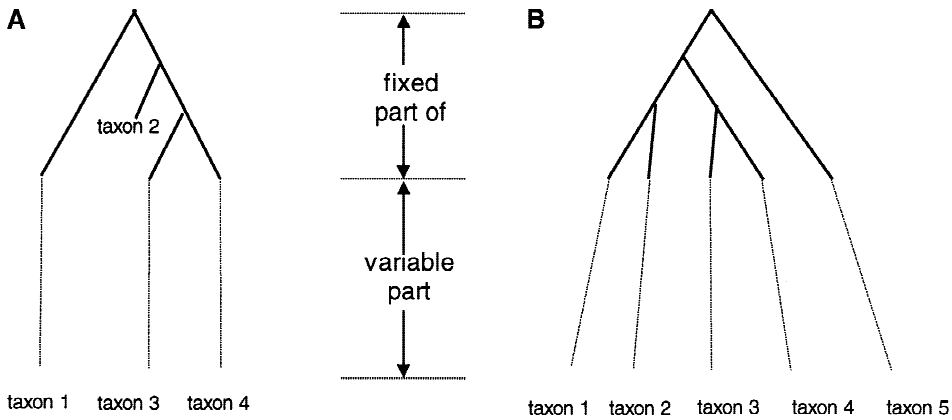

Fig. 6. The trees used in simulations of the covarion model: (**A**) four-taxon; (**B**) five-taxon. The lengths of internal edges of the trees were held constant, and the external edges (except for the edge to taxon 2 in A) ranged over different lengths (times) for successive simulations. Data were simulated on these trees using the parameters in Fig. 5, and then trees inferred from the data using Phylip with a Kimura 2ST model. Results for the five-taxon tree are shown in Fig. 7.

do not even use using measured rates of evolution. Consequently, it is impossible to apply any conclusions to real data and real times of divergence. In many ways, the conclusion in Fig. 2 is self-evident; current methods should fail for ancient divergences under the standard assumptions of the models. It is not scientific just to assume that models will work for the oldest divergences; they must be tested explicitly.

Before examining ways in which covarion-type models may help ancient divergences, it is necessary to be cautious of existing knowledge about deep divergences in the tree of life. There is no good biochemical reason for current models to be reliable for inferring the early branches of the tree of life. It was for this reason that we used relics of the RNA world as a possible alternative for rooting the tree of life (Poole et al. 1998). Given the difficulties in inferring the order of divergence of mammals correctly, it is certainly premature to be confident in ancient divergences. However, the caution works both ways; just because two genes give different trees does not mean that one (or both!) have been subject to lateral transfer. It is expected that genes can differ in the trees they predict; they do even for mammals (Penny et al. 1982). Lateral transfer is undoubtedly an important feature in evolution, but the results presented here caution against invoking it every time two genes differ in the tree they predict.

The present results (Fig. 7) are consistent with our working hypothesis—the covarion model predicts that trees for ancient divergences will be better than expected from simple biochemical models. In a sense, the covarion model increases the "effective number" of variable sites. The covarion model could also explain why a particular molecule might have a range of times for which it is most suitable (e.g., Graybeal 1994; Whitfield and Cameron 1998). This is because the length of time it takes a particular protein to saturate depends on the rate of evolution of its tertiary structure. If the tertiary structure does not change, the protein is expected to saturate sooner (see Griffiths 1997). Other authors (e.g., Simon et al. 1994) suggest that, in practice, some macromolecules lose resolution (as expected) at intermediate dates of divergence but improved again for divergences that were even older. Such a result could occur if some slight changes to secondary and tertiary structure occurred only very occasionally (that is, low values of δ, or no longer a stationary model). In such circumstances, new invariant positions that helped recovery of the tree would arise occasionally. An alternative may be an occasional, but larger, change in covarion structure (see Lockhart et al. 1996; Wolfe and dePamphilis 1998). For ancient divergences, individual proteins might not allow resolution because their 2-D and 3-D structure is too highly conserved. It is possible too that, on average, ribosomal RNA does better than expected because its secondary structure does vary considerably. We have recently shown that RNA secondary structure itself can be used to reconstruct trees, even when the sequences cannot be aligned with any confidence (Collins et al. 2000).

Some of the most interesting recent applications of a covarion model are studies by Brinkmann and Philippe (1999) and Lopez et al. (1999). They use concepts from the covarion model to identify the slowest-evolving sites. This differs from the RAS approach, which does not identify which sites are faster or slower, just the distribution of rates. Lopez et al. report that the slower- and faster-evolving sites support different trees! Thus the faster sites appear to be more strongly affected by the long-edges-attract problem (Hendy and Penny 1989).

Although the covarion model in general could be considered "good news," there are times when a nonstationary version of the covarion model could be "positively misleading." In these cases, a covarion process could reinforce support for an incorrect tree, and a possible example has been discussed (Lockhart et al. 1996, 1998; Steel et al. 2000). One case has a tree with five major

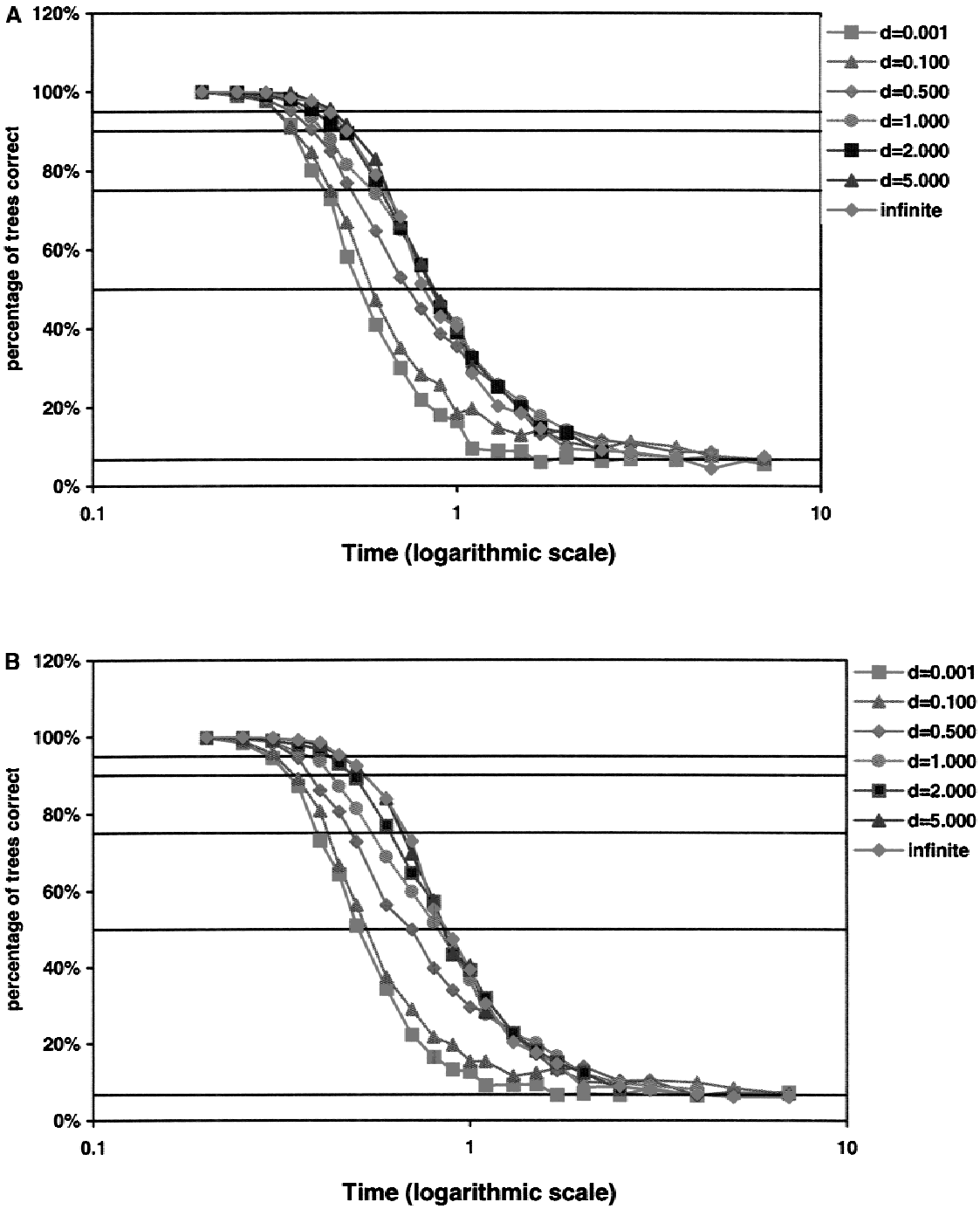

Fig. 7. Results for neighbor-joining (**A**) and maximum likelihood (**B**) for data simulated on the tree in Fig. 6B and under a covarion model. Each point represents 1000 simulations for sequences of 1000 nucleotides. The *x*-axis shows the relative times on a logarithmic scale, and the *y*-axis the probability of recovering the generating tree correctly. In each figure, the seven curves are δ (*d*) values—the rate of interconversion between fixed and variable sites. Increasing δ enhances the chance of recovering the correct tree, allowing the correct tree to be inferred for 50–100% longer divergence times.

branches, each with many sequences (Lockhart et al. 1996). On two branches, there have been major (but different) changes in the function of the gene, and the covarion set has changed independently in these two groups. The consequence is that branches where genes retain their initial function tend to group together, even though they may not have been adjacent on the original tree. That example of a covarion model differs from the stationary version presented here in that it has occasional, but large, changes in the covarion set; it is a nonstationary model. In our implementation, the continued small covarion changes are independent of position on the tree; the process does not change across the tree. The Lockhart et al. results emphasize again the need to

understand the changes that are occurring in a gene during evolution.

The results in Figs. 2 and 7 are for ideal cases: no changes in nucleotide composition, no positive selection for similar changes on different lineages, no correlation between sites, no lateral transfer, etc. These additional factors are expected to make it even more difficult to recover trees accurately from real data. For example, it has recently been shown that some data sets (Cao et al. 1998) show contradictions between different proteins in the mitochondrial genome. This shows that there are different signals in the data, and even longer sequences may be required than lengths estimated from simulations. Nevertheless, the results (Fig. 7) are consistent with the original hypothesis that the covarion model gives a biochemical basis for why tree reconstruction may, in principle, do better for ancient divergences than simple biochemical models predict.

Relationships Between Models

One aspect of our model makes it more general than that of Fitch and Ayala (1994)—the interconversion between fixed and variable states is continuous in our formulation. That is, they are not limited to a change in the sequence of the macromolecule. Our formulation is simpler to analyze because the two processes (changes between nucleotides and interconversion between fixed and variable states) are independent. This additional flexibility is realistic. The set of variable sites may be altered by both intramolecular and intermolecular interactions (Lockless and Ranganathan 1999), as well as by environmental changes such as in temperature. Although either formulation (Fitch and Ayala's, or ours) can be justified biologically, the calculation and analysis are more straightforward for our hidden Markov process.

Our formulation of the covarion model can be considered within an expanded class of i.i.d models (see Penny et al. 1992). Changes are independent, both between sites and on different lineages of the tree, and in addition, each site is drawn from the same (identically distributed) distribution. Models more general than the Kimura 3ST can be implemented under a covarion model, but then the Hadamard matrix cannot be used for diagonalizing matrices—though other methods are available. We consider our present model a member of the Kimura–Dickerson–Fitch class of models. It includes the following elements: it is a stochastic model; most changes are neutral; a restricted number of sites is free to vary at any one time; but that set changes with time.

It is interesting to note that the covarion model gives some biochemical justification for the use of, for example, a Γ distribution of rates. Operationally, the Γ distribution compensates, in part, for some sites being invariant (Waddell and Steel 1997). Theoretically, Tuffley and Steel (1997) report that for *pairs* of sequences, a covarion model can always be matched by a Γ (or more general) distribution. This does not generalize to higher numbers of sequences, and tests are available that can distinguish RAS and covarion models (Tuffley and Steel 1997; Lockhart et al. 1998, 2000). The tests compare the numbers of constant (or varied) sites in different parts of the tree, but little is known of the power of the tests. Thus the equivalence of covarion and RAS models is only for pairs of taxa, and further work is required to determine when the Γ distribution is, in practice, a useful approximation to a covarion model. At present, we are more interested in exploring the usefulness of identifying faster and slower sites (see Brinkmann and Philippe 1999; Lopez et al. 1999), rather than assuming that a site is sampled from a distribution of RAS. Finally, the covarion model is also an explanation for the common practice of discarding sites that are difficult to align. Such difficult-to-align sites are expected to occur where there have been changes in the 3-D structure of the macromolecule.

Role of Mathematical Models

This relationship between the covarion model and an approximation to it by a Γ distribution (Tuffley and Steel 1997) raises another interesting question: the role in science of formal mathematical descriptions and underlying physical models. The most mathematically developed aspects of biology include population genetics, ecology, physiology, and biochemical kinetics. In each case, the mathematical model is a formalization of the underlying biological mechanisms. However, with RAS models there has been little attempt to consider the biochemistry that might "justify" a Γ (or other) distribution. In general, scientists prefer the mathematics to be based on a physical (biological) model, not just be an arbitrary mathematical description. Should this also be the case in evolutionary analysis?

There is one established viewpoint, instrumentalism, that accepts that mathematical models may be useful "instruments" for calculation.

> There is no need for these hypotheses to be true, or even to be at all like the truth; rather, one thing is sufficient for them—that they should yield calculations which agree with the observations. [See discussion by Popper (1963, pp. 97ff)].

The best-known case in science is the opportunity given to Galileo to use this reasoning as an alternative to the heliocentric hypothesis of the planets orbiting the sun. The church hierarchy accepted that Galileo's calculations were more accurate—it would be acceptable just to deny that the equations described the solar system. Sober

(1998) discusses some aspects of modern science where instrumentalism is still influential, and a case can be made for some statistical models that successfully identify a pattern, without identifying the underlying processes.

In the case of the distribution across sites model, some authors may be satisfied if the distribution aids in getting the correct tree, regardless of any biochemical process that may, or may not, underlie the calculation. We prefer, however, that more consideration be given to the relationship between mathematical models and the underlying biochemical mechanisms. It is little more than a tautology just to say, "Sites evolve at different rates," without understanding the mechanisms involved. It is more satisfying to have a biochemical mechanism described mathematically, rather than a convenient mathematical description not based on any biological mechanism. It is interesting that more biochemical realism is being introduced into models of molecular evolution; examples include Goldman and Yang (1994), Liò and Goldman (1999), Thorne et al. (1992), and Schöniger and von Haeseler (1994).

Extensions to the present hidden Markov implementation of the covarion model are straightforward. The present version already allows a proportion of sites to be permanently fixed or permanently variable. A RAS model could be included so that the variable sites evolve at different rates (though a biochemical explanation is unclear). It is straightforward to add another layer of invariable sites with only one class of invariant sites able to become covarions directly (that is, invariant 2 $\leftrightarrow$ invariant 1 $\leftrightarrow$ covarions). Other sites may always be fixed; others may always be variable and will saturate relatively quickly. In such cases, it would be necessary to detect the slower-evolving sites to study ancient divergences (Brinkman and Philippe 1999; Lopez et al. 1999). Another extension is a maximum-likelihood implementation of the covarion model, including estimating the optimal values for φ and δ (A. Rambaut, personal communication). Hidden Markov models have been used in other aspects of recovering evolutionary information (Baldi et al. 1994; Felsenstein and Churchill 1996; Krogh et al. 1994). Such models are still relatively underexplored in molecular evolution and will probably turn out to be as useful here as in many other areas of science.

It is premature to decide how useful our covarion-like model will be in practice. The present paper is just one contribution focusing on the underlying molecular biology of molecular evolution models. There are serious problems in studying ancient divergences, both theoretical and practical. We think that a reasonable case has been made to take the covarion model seriously, and there may be other ways of including basic molecular biology knowledge into evolutionary models. Future work requires an improved synthesis of mathematics and biochemical realism.

Acknowledgments. We thank Ted Drawneek for assistance with the parallel use of multiple computers, P.J. Lockhart and two anonymous referees for many useful comments, and the New Zealand Marsden Fund for financial support.

Note Added at Proof

Galtier (2001) has recently described a maximum-likelihood implementation of the covarion-like model described here.

References

Baldi P, Chauvin Y, Hunkapiller T, McClure MA (1994) Hidden Markov models of biological primary sequences information. Proc Natl Acad Sci USA 91:1059–1063

Brinkmann H, Philippe H (1999) Archaea sister-group of bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol Biol Evol 16:817–825

Camardella L, Caruso C, D'Avino R, Di Prisco G, Rutigliano B, Tamburrini M, Fermi G, Perutz MF (1992) Haemoglobin of the Antarctic fish *Pagothenia bernacchii.* Amino acid sequence, oxygen equilibria and crystal structure of its carbonmonoxy derivative. J Mol Biol 224:449–460

Cao Y, Janke A, Waddell PJ, Westerman M, Takenaka O, Murata S, Okada N, Pääbo S, Hasegawa M (1998) Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders. J Mol Evol 47:307–322

Carugo O, Argos P (1997) NADP-dependent enzymes II: Evolution of the mono- and dinucleotide binding domains. Proteins Struct Funct Gen 28:29–40

Chang J (1996) Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. Mol Biosci 134:189–215

Charleston MA, Hendy MD, Penny D (1994) The effects of sequence length, tree topology and number of taxa on the performance of phylogenetic methods. J Comput Biol 1:133–151

Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5:823–826

Chothia C, Lesk AM (1987) The evolution of protein structures. Cold Spring Harb Symp Quant Biol 52:399–405

Collins LJ, Moulton V, Penny D (2000) Use of RNA secondary structure for evolutionary relationships: The case of RNase P and RNase MRP. J Mol Evol 51:194–204

Dickerson RE (1971) The structure of cytochrome c and rates of molecular evolution. J Mol Evol 1:26–45

Elliot RJ, Aggoun L, Moore JB (1995) Hidden Markov methods: Estimation and control. Springer-Verlag, New York

Felsenstein J (1997) http://evolution.genetics.washington.edu/phylip/software.html

Felsenstein J, Churchill GA (1996) A hidden Markov model approach to variation among sites in rate of evolution. Mol Biol Evol 13:93–104

Fisher A, Shi Y, Ritter A, Ferretti JA, Perez-Lamboy G, Shah M, Shiloach J, Taniuchi H (2000) Functional correlation in amino acid residue mutations of yeast iso-2-cytochrome c that is consistent with the prediction of the concomitably variable codon theory in cytochrome c evolution. Biochem Gen 38:181–200

Fitch WM (1971) Rate of change of concomitantly variable codons. J Mol Evol 1:84–96

Fitch WM (1986) The estimate of total nucleotide substitutions from pairwise differences is biased. Phil Trans R Soc Lond B 312:317–324

Fitch WM, Ayala FJ (1994) The superoxide dismutase molecular clock revisited. Proc Natl Acad Sci USA 91:6802–6807

Fitch WM, Markowitz E (1970) An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem Genet 4:579–593

Galtier N (2001) Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol Biol Evol 18:866–873

Gillespie JH (1988) More on the overdispersed molecular clock. Genetics 118:85–386

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11: 375–736

Graybeal A (1994) Evaluating the phylogenetic utility of genes: A search for genes informative about deep divergences among vertebrates. Syst Biol 43:174–193

Griffiths CS (1997) Correlation of functional domains and rates of nucleotide substitution in cytochrome b. Mol Phylogenet Evol 7: 352–365

Hendy MD, Charleston MA (1993) Hadamard conjugation: A versatile tool for modelling sequence evolution. NZ J Bot 31:231–237

Hendy MD, Penny D (1989) A framework for the quantitative study of evolutionary trees. Syst Zool 38:297–309

Hendy MD, Penny D, Steel MA (1994) Discrete Fourier spectral analysis of evolution. Proc Natl Acad Sci USA 91:3339–3343

Karon JM (1979) The covarion model for the evolution of proteins: Parameter estimates and comparison with Holmquist, Cantor, and Jukes' stochastic model. J Mol Evol 12:197–218

Kimura M (1981) Estimation of evolutionary distance between homologous nucleotide sequences. Proc Natl Acad Sci USA 78:454–458

Koonin EV, Gorbalenya AE (1989) Evolution of RNA genomes: Does the high mutation rate necessitate high rate of evolution of viral proteins? J Mol Evol 28:524–527

Krogh A, Brown M, Mian IS, Sjölander K, Haussler D (1994) Hidden Markov models in computational biology: Applications to protein modeling. J Mol Biol 235:1501–1531

Lesk AM (2000) Introduction to protein architecture: The structural biology of proteins. Oxford University Press, Oxford

Levitt M, Gerstein M (1998) A unified statistical framework for sequences comparison and structural comparison. Proc Natl Acad Sci USA 95:5913–5920

Li W-H (1997) Molecular evolution. Sinauer Associates, Sunderland, MA, p 75

Liò P, Goldman N (1999) Using structural information in evolutionary inference: Transmembrane proteins. Mol Biol Evol 16:1696–1710

Lockhart PJ, Larkum AWD, Steel MA, Waddell PJ, Penny D (1996) Evolution of chlorophyll and bacteriochlorophyll: The problem of invariant sites in sequence analysis. Proc Natl Acad Sci USA 93: 1930–1934

Lockhart PJ, Steel MA, Barbrook AC, Huson D, Charleston MA, Howe CJ (1998) A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. Mol Biol Evol 15: 1183–1188

Lockhart PJ, Huson D, Maier U, Fraunholz MJ, Van de Peer Y, Barbrook AC, Howe CJ, Steel MA (2000) How molecules evolve in bacteria. Mol Biol Evol 17:835–838

Lockless SW, Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. Science 286: 295–299

Lopez P, Forterre P, Philippe H (1999) A method for extracting ancient phylogenetic signal: The rooting of the universal tree of life based on elongation factors. J Mol Evol 49:496–508

Marshall CE, Raff EC, Raff RA (1994) Dollo's law and the death and resurrection of genes. Proc Natl Acad Sci USA 91:12283–12287

Miyamoto MM, Fitch WM (1995) Testing the covarion hypothesis of molecular evolution. Mol Biol Evol 12:503–513

Ohta T, Kimura M (1971) On the constancy of the evolutionary rate of cistrons. J Mol Evol 1:18–25

Page RDM, Holmes EC (1998) Molecular evolution: A phylogenetic approach. Blackwell Science, Oxford

Palumbi SR (1989) Rates of molecular evolution and the fraction of nucleotide positions free to vary. J Mol Evol 29:180–187

Pascarella S, Argos P (1992) Analysis of insertions/deletions in protein structures. J Mol Biol 224:461–471

Penny D (1993) Mathematics yes, physics no. Nature 366:504

Penny D, Foulds LR, Hendy MD (1982) Testing the theory of evolution by comparing phylogenetic trees constructed from five different protein sequences. Nature 297:197–200

Penny D, Hendy MD, Steel MA (1992) Progress with evolutionary trees. Trends Ecol Evol 7:73–79

Poole AM, Jeffares DC, Penny D (1998) The path from the RNA world. J Mol Evol 46:1–17

Popper KR (1963) Conjectures and refutations: The growth of scientific knowledge. Routledge and Kegan Paul, London

Renault L, Nassar N, Vetter I, Becker J, Klebe C, Roth M, Wittinghofer A (1998) The 1.7 A crystal structure of the regulator of chromosome condensation (RCC1) reveals a seven-bladed propeller. Nature 392:97–101

Schöniger M, von Haeseler A (1994) A stochastic model for the evolution of autocorrelated DNA sequences. Mol Phylogenet Evol 3: 240–247

Simon C, Frati F, Beckenbach A, Crespi B, Liu H, Flook P (1994) Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation conserved PCR primers. Ann Entomol Soc Am 87:651–701

Sober E (1998) Instrumentalism revisited. Critica 31:3–38

Spiller B, Gershenson A, Arnold FH, Stevens RC (1999) A structural view of evolutionary divergence. Proc Natl Acad Sci USA 96: 12305–12310

Steel MA, Penny D (2000) Parsimony, likelihood and the role of models in molecular phylogenetics. Mol Biol Evol 17:839–850

Steel MA, Huson D, Lockhart PJ (2000) Invariable site models and their use in phylogeny reconstruction. Syst Biol 49:225–232

Steel MA, Székely LA, Hendy MD (1994) Reconstructing trees when sequence sites evolve at different rates. J Comp Biol 1:153–163

Thorne J, Kishino H, Felsenstein J (1992) Inching toward reality: An improved likelihood of sequence evolution. J Mol Evol 34:3–16

Tuffley C, Steel MA (1997) Modeling the covarion hypothesis of nucleotide substitution. Math BioSci 147:63–91

Uzzell T, Corbin KW (1971) Fitting discrete probability distributions to evolutionary events. Science 172:1089–1096

Van de Peer Y, Rensing SA, Maier U-G, DeWachter R (1996) Substitution rate calibration of small subunit ribosomal subunit RNA identifies Chlorarachnida nucleomorphs as remnants of green algae. Proc Natl Acad Sci USA 93:7732–7736

Waddell PJ, Steel MA (1997) General time-reversible distances with unequal rates across sites: Mixing Γ and inverse Gaussian distributions with invariant sites. Mol Phylogenet Evol 8:398–414

Waddell PJ, Penny D, Hendy MD, Arnold GC (1994) Variance-covariances matrices for evolutionary trees using Hadamard transforms. Mol Biol Evol 11:630–642

Waddell PJ, Penny D, Moore T (1997) Hadamard conjugations and modeling sequence evolution with unequal rates across sites. Mol Phylogenet Evol 8:33–50

Waddell PJ, Cao Y, Hauf J, Hasegawa M (1999) Using novel phylogenetic methods to evaluate mammalian mtDNA. Syst Biol 48:31–53

Whitfield JB, Cameron SA (1998) Hierarchical analysis of variation on the mitochondrial 16S rRNA gene among Hymenoptera. Mol Biol Evol 15:1728–1743

Wolfe AD, dePamphilis CW (1998) The effect of relaxed functional constraints on the photosynthetic gene rbcL in photosynthetic and nonphotosynthetic parasitic plants. Mol Biol Evol 15:1243–1258

Yang Z (1996) Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol Evol 11:367–372

Zuckerkandl E (1976) Evolutionary processes and evolutionary noise at the molecular level. J Mol Evol 7:269–311