

Overview:

Mathematical elegance with biochemical realism: the covarion model of molecular evolution

David Penny¹, Bennet J McComish¹, Michael A Charleston^{2,3} and Michael D Hendy²

¹ Institute of Molecular BioSciences, and

² Institute of Fundamental Sciences, Massey University, Palmerston North, New Zealand

³ current address, Zoology Dept, University of Oxford

Address for correspondence:

David Penny
Institute for Molecular BioSciences
P O Box 11222
Massey University,
Palmerston North
New Zealand.
e-mail d.penny@massey.ac.nz
Courier address, Science Towers D5.01

Tel +64 6 350 5033
Fax: +64 6 350 5694

Key words: covarion model, hidden Markov model, molecular evolution, rates across sites, role of models in science.

ABSTRACT

There is an apparent paradox in our understanding of molecular evolution. *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 gamma distribution of rates-across-sites -RAS) may allow 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, which makes justification for the rate-constancy assumption of purely mathematical models problematic. 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 suggest that such an approach provides a better framework for studying anciently-diverged sequences than the standard rates-across-sites models. 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.

The earliest models for the evolution of sequence data assumed that all variable sites evolved at the same rate. Two approaches were introduced early in the study of molecular evolution 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 of Fitch and Markovitz (1970) and 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 gamma distribution (see Yang et al. 1996), has been well developed mathematically but several other distributions (for example, 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 only require one or two additional parameters, irrespective of the length of the sequences. However a potential problem with these rates-across-sites models is that a biochemical explanation has not been developed as to how sites maintain the same potential rate over long evolutionary periods. To illustrate this problem, some features of the structural evolution of proteins are considered next.

Structural evolution of proteins

It is difficult to find a biochemical mechanism that would maintain the same potential rate of evolution at a site - irrespective of whether the gene was in eukaryotes, archaea or eubacteria, or within thermophiles or mesophiles. In fact, biochemical information predicts the opposite; sites in widely different lineages should vary in their rates. The following examples show that one of the strongest conclusions of structural biology is that the 3-D structure of orthologous proteins does vary during evolution. This has been demonstrated, for example, by changes in the root-mean-square (rms) difference in the position of the alpha carbon atoms (C_{α}) along the backbone of the 3-D structure of a protein, and is measured in Angstroms (\AA). 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 sequence divergence - even if only considering the core of the proteins. The effect was non-linear, with increasing difference in 3-D structure at higher sequence divergence. Similar effects were found with 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 are also 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\AA , though the closeness of the match varies throughout the protein (Camardella et al. 1992, for example their Fig

8). Another example are 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, their 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. A recent overview of protein structural evolution is Lesk (2000, chapters 5 and 6).

The over-riding 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 (rates-across-sites) approach has desirable mathematical properties the requirement for each site to maintain its characteristic rate throughout evolution is not in agreement with our current understanding of biochemical processes.

Covarion model

The covarion substitution model (Fitch and Markovitz 1970; Fitch 1971) was introduced at the same time as the rates-across-sites 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, but fixed in others. This switch would result from slight changes during evolution in secondary and tertiary structure that were 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:

“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*. ” (p585)

‘Covarion’ is a contraction of *concomitantly variable codons*, 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 covarion for the nucleotide version; the underlying concepts are identical irrespective of the number of character states and it is 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 (for example, Dickerson 1971). (Here we are considering differences in the rate of evolution between proteins, not differences between sites within a protein.) We refer to these earliest explanations of differential rates in protein evolution as the Kimura-Dickerson model. Under this basic biochemical model changes in sequences were considered stochastic and neutral (Kimura), and variation in the proportion of unconstrained sites which were 'free to vary' accounted for the different rates between proteins (Dickerson). Several authors 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 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 change in 3-D structure (and consequently of structural constraints) of a protein. (Remembering that under this model the slower average rate at 1st and 2nd positions is because fewer sites are viable.) 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 model where some sites switch between 'on' and 'off' states 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-establish covarion name, rather it is to show that the model is made up of components, thus making it easier to analyse. Karon (1979), using Fitch's cytochrome c data, improved the original model to more fully account 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 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 later. 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 gamma distribution of rates-across- sites. Similarly, a new quantitative test (Lockhart et al. 1998) shows that for their data a covarion model fits the data better than a rates-across-sites model - or more accurately, the test rejects any RAS model where each site is always has the same rate. Several other authors have limited discussions of

the covarion model (for example, 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 appears to have 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, it has been proven (Steel et al. 1994) 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 (rates-across-sites) approach, the covarion model has a reasonable biochemical foundation, but has appeared to lack the required mathematical properties. However, before proceeding with an implementation the covarion model, we will 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 that assumes a site is always in the same rate class would lead to sequences saturating relatively quickly during evolution. This is easily studied by simulation. For example, if sites evolve at 0.5% per million years (Li 1997 p75; Page and Holmes 1997 p239), then there will be an average of one change per site after 200 million years. From previous computer simulations, recovering evolutionary trees is inaccurate when there has been, on average, more than about one change per site (Charleston et al. 1994).

To illustrate this problem for ancient divergences and simple biochemical models (that is, neither covarion nor rates-across-sites models) simulations were run on 300,000 randomly selected 9-taxon trees. The substitution rate was increased logarithmically so that the expected numbers of changes per site (which is linear with time) increased from 0.025 to 10 (equivalent to 5 to 2000 million years of evolution). 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 2-state characters using the Hadamard conjugation (Hendy and Charleston 1993). Data was generated for 100-1000 variable sites (that is, excluding any invariable sites). Trees were inferred from each data sample using neighbor-joining (Felsenstein 1997) - which is consistent under this model.

The results in figure 2 show the expected error in recovering the correct tree for nine taxa (six internal edges), and are for 1000 variable sites. The increasing time of divergence (x-axis) is equivalent to 5 to 2000 million years of evolution. The y-axis is the proportion of times trees with 0, 1, ..., 6 internal edges (branches) are inferred correctly - six implying that the correct tree is fully

recovered. 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 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 this 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.

The conclusion is that under simple biochemical models (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. Although in practice the long edges on the tree may be broken up by other taxa, the example illustrates the problem of current models for ancient divergences. Under these standard models, there is no justification yet 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 their 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 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 Figure 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 with different data sets (both molecular and morphological). Bird sequences don't come out among mammals or invertebrates, mosses among flowering plants or fungi, and so on. There are many difficulties with differences between data sets, and with the oldest divergences, but there is no suggestion (as in Figure 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 just being resolved. The covarion model offers a possible resolution to this fundamental problem of molecular evolution. But first other alternatives should be considered that still assume a site is always in the same rate class.

Alternative models

A standard answer may be 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. How would a site in a protein ‘change more slowly’? Possibilities include differences in mutation rates and/or in selective constraints. A difference in mutation rate between codon positions in a single gene is not considered likely as a general explanation, and we now show that differences in selective constraints do not work either. With the basic Kimura-Dickerson model of molecular evolution it is assumed that from 1-4 nucleotides are viable at a site. Conversely, negative selection would eliminate from 0 to 3 mutations at that site.

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 some sites evolved more slowly because only two nucleotides (or amino acids) were viable at the site. The rate of loss of phylogenetic signal was calculated on a 4-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 were then made longer and longer, and the results shown as the x-axis in Figure 3. Time is on a logarithmic scale for 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 either two, or with four, nucleotides viable at a site. Conversely, ‘wrong_2’ and ‘wrong_4’ are the probabilities of a site supporting either 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 $t = \text{four taxa}$) only eight 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 nucleotides viable (one out of eight $[2^{t-1}]$ patterns) and to 0.0469 for four nucleotides (three patterns out of 64 $[4^{t-1}]$).

In Figure 3 there is little difference in the rate of randomization (loss of phylogenetic signal) with either two or four nucleotides viable at a site. Under the parameters used, sites with either two or four nucleotides viable are expected to be misleading by 200 million years. By 300-400 million years, they are approaching randomization. With only two nucleotides viable, there is a 1-3% slower approach to randomization - essentially no difference. Another conclusion from the results in Figure 3 [not shown] was that the proportion of sites that were free to vary but had not changed, decreased to zero relatively quickly. This reinforces the conclusion that sites that are constant for anciently diverged trees are functionally constrained (invariable). Such sites should not be used when estimating either the number of multiple changes at other sites, or the nucleotide compositions of 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). Overall, the results in figure 3 reinforce the conclusion that the basic models do not give a

biochemical justification for why sequences are effective at reconstructing history for ancient divergences.

An alternative explanation for some sites evolving more slowly is that some slightly deleterious mutations are occasionally fixed in a population. These sites would change more slowly than the neutral rate (because some mutations are eliminated by negative selection). By itself, this is not a general solution because it implies a slow continual decline in fitness over time. A different approach is due to Zuckerkandl (1976) where positive selection was invoked for fixation of virtually all changes. On 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 lead 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 changes were neutral. Indeed, positive selection for similar changes on different lineages would only make it harder to recover the correct tree. In general, simple models where a site always has the same number of nucleotides that are viable, while other mutations are eliminated, do not offer a biochemical explanation for sites differing markedly in their rates of evolution.

The results in Figures 2 and 3 illustrate a fundamental problem with inferring ancient evolutionary trees from sequence data; current biochemically-based models are not encouraging about 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 was generated under a simple covarion model and then analyzed by standard (non-covarion) models.

Covarion model with two additional parameters.

Covarion models, where sites vary in their rate of evolution as the 3-D structure of the protein or RNA evolves, 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 any model that predicts sites are always in the same rate class. This present paper gives the basic reasoning 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 3ST model (Kimura 1981) - which contains the 2ST and Jukes-Cantor models as special cases. The second (hidden) process has the two additional parameters, the proportion of sites (φ) that are free to vary (the covarions), and the rate of interchange (δ) between these variable states and sites that cannot vary. All sites have the same chance 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 is unchanged over the whole tree. At a variable site, a mutation may be fixed in the population by either 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 particular site in a sequence is able to change is unknown - (hence the name 'hidden' Markov model). The variability status is represented by a superscript: plus '+' when the states are free to change (A^+ , G^+ , C^+ , & T^+) and minus '-' when fixed (A^- , G^- , C^- , & T^-). For proteins, the forty 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 variable states is set to maintain the proportion φ of variable (covarion) and fixed sites.

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

Kimura's 3ST model is used by both RAS and covarion models and is described by an instantaneous rate matrix, K , which is diagonalized by the Hadamard matrix H :

$$K = \begin{matrix} A^+ \\ G^+ \\ C^+ \\ T^+ \end{matrix} \begin{bmatrix} * & \alpha & \beta & \gamma \\ \alpha & * & \gamma & \beta \\ \beta & \gamma & * & \alpha \\ \gamma & \beta & \alpha & * \end{bmatrix} \quad H = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \quad \Lambda = H^{-1}KH = \begin{bmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 \\ 0 & 0 & 0 & \lambda_4 \end{bmatrix},$$

where ‘*’ = $-(\alpha + \beta + \gamma)$ (in order that the rows sum to zero); $H^{-1} = \frac{1}{4}H$ 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 , $\exp(K) = I + K + K^2/2! + K^3/3! + K^4/4! + \dots = H \cdot \exp(\Lambda) H^{-1}$, and $\exp(\Lambda)$ is the diagonal matrix whose entries are $\exp(\lambda_i)$. The transition matrix $M = \exp(K)t$ 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 a simulation.

A general rate matrix (M) for our hidden Markov model is,

$$K' = \begin{bmatrix} K & 0 \\ 0 & 0 \end{bmatrix} + \alpha \begin{bmatrix} -I & I \\ I & -I \end{bmatrix} \text{ where } K' \text{ is an } 8 \times 8 \text{ matrix and } K \text{ and } I \text{ are } 4 \times 4 \text{ matrices.}$$

The instantaneous rate matrix (K') for the hidden Markov model we use is shown in Figure 5.

Our working hypothesis was that the covarion model would allow older divergences to be recovered more accurately. The trees chosen to test the covarion model (figure 6) are inconsistent with uncorrected parsimony (Hendy and Penny, 1989) and are especially difficult to recover without accurate corrections for multiple changes. The four taxon tree $(t_1, (t_2, (t_3, t_4)))$ (figure 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 5-taxon tree $((t_1, t_2), (t_3, t_4), t_5)$ all lineages fitted the molecular clock (figure 6B). However, with uncorrected parsimony the correct tree (even with infinitely long sequences) is the 12th longest (out 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 interest was whether the correct tree was recovered correctly as the total amount of evolution increased. To test this, 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 $\beta = 0.0025$ (the transition rate) and with equal nucleotide frequencies at the root. The two additional parameters for the covarion model, ϕ and δ , were 0.5 for ϕ (the proportion of variable sites) and varied for δ (the rate of interchange between fixed and variable). 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 5-taxon tree is large. Consequently the simulation and tree-building steps were performed on fifty 133MHz 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 5-taxon tree (figure 6B) are shown in figure 7 (the 4-taxon tree in figure 6A has similar results). The x-axis is time (on a logarithmic scale) and the y-axis the probability of recovering correctly the generating tree. Figure 7A is for neighbor-joining and 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 figure 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 1, 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 may perform better than reported in figure 2.

The extent of the improved performance, under these conditions, is shown by the observation that the model tree can still be identified after longer periods of evolution. 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 4-taxon tree, but with the 5-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 in order to test predictions more constructively.

Consequences for ancient divergences

It is clear that simulation studies need to use both realistic times and real rates of evolution, and not average rates across invariable sites. Many simulation studies use only relatively short evolutionary periods without using realistic rates of evolution. Consequently, it is impossible to convert their conclusions to real data and real times of divergence. In many ways, the conclusion in Figure 2 is self-evident, current methods should fail for ancient divergences under the standard assumptions behind 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 basis 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 inferring correctly the order of divergence of mammals, 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 disagree.

The present results (Figure 7) are consistent with our working hypothesis - the covarion model may explain why trees are often better than expected under the biochemical models currently in use. 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 (for example, 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 tertiary structure does not change, the protein is expected to saturate sooner (see Griffiths 1997). Others (for example, Simon et al. 1994) suggest that, in practice, some macromolecules lost 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 only occurred very occasionally (that is, low values of δ , or no longer a stationary model). Under 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 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 rates-across-sites approach, which does not identify which sites are faster or slower, just the distribution of rates. Lopez et al. reports that the slower and faster evolving sites support different trees! The faster sites appear 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 non-stationary version of 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 branches, each with many sequences (Lockhart et al. 1996). On two branches, there have been major (but different) changes in 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 although 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 non-stationary 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 for figures 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 estimated from simulations. However, the results (Figure 7) are consistent with the original hypothesis that the covarion model gives a biochemical basis for how tree reconstruction may, in principle, do better than simple biochemical models predict for ancient divergences.

Relationships between models

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

Our formulation of the covarion model can be considered within an expanded class of i.i.d models (independent and identically distributed, 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 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 a Kimura-Dickerson-Fitch class of models. It includes the elements of a stochastic model, most changes neutral, with differences in numbers of sites free to vary at any one time, but that set changing with time.

It is interesting to note that the covarion model gives some biochemical justification for the use of, for example, a gamma distribution of rates. The gamma distribution compensates, in part, for some sites being invariant (Waddell and Steel 1997). In addition, Tuffley and Steel (1997) report that for pairs of sequences, a covarion model can be matched by a gamma (or a more general) distribution. Tests are now available for distinguishing between a rates-across-sites model, and a covarion model (Tuffley and Steel 1997; Lockhart et al. 1998; 2000). The tests work by comparing numbers of constant (or varied) sites in different parts of the tree. They show that a covarion model is a good description of the data, and that conversely rates-across-sites models (where sites are always in the same rate category) are not valid for the data sets tested. Work is required to determine when the gamma distribution is a useful approximation to the 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 a site is sampled from a distribution of rates-across-sites. Finally, the covarion model is perhaps a justification for the common practice of discarding sites that are difficult to align. Such difficult to align sites are expected to occur where there has been a change in 3-D structure of the macromolecule.

The role of mathematical models

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

There is however one established viewpoint, instrumentalism, that accepts that mathematical models may be useful 'instruments' for calculation. The best known case in science is the opportunity given to Galileo to use this reasoning in relation to the heliocentric hypothesis of the planets orbiting the sun

“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 in Popper 1963 pp 97ff).

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 to be given in molecular evolution on 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 preferable to have a biochemical mechanism that can then be described mathematically, rather than just have a convenient mathematical description of the data 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; Thorne et al. 1992; Schöniger and von Haeseler 1994).

Extensions to the present hidden Markov implementation of the covarion model are possible. The model already allows a proportion of sites to be permanently fixed. A ‘rates-across-sites’ 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 (that is, invariant2 \leftrightarrow invariant1 \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 (Brinkmann 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 pers. comm.). 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 under-explored 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 to focus on the underlying molecular biology of models of molecular evolution. There are serious problems studying ancient divergences, both theoretical and practical. We think that a reasonable case has been made to take the covarion model seriously, though there may be other ways of including basic molecular biology knowledge into evolutionary models. Future work requires improved synthesis mathematics and biochemical realism.

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

LITERATURE CITED

- Baldi P, Chauvin Y, Hunkapiller T, McClure MA (1994) Hidden Markov models of biological primary sequence 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* 224:449-60
- 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.* Subject to editorial revision).
- 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
- 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 Gen* 4:579-593

- 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 Phylog 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 Univ. Press, Oxford (in press)
- 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, p75
- 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: (in press)
- 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 covarion model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. *Mol Biol Evol* 15:1183-1188
- 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 M.D (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 Å 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 Phyl 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. *Annals Entomol Soc Am* 87:651-701
- 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.* (accepted, subject to editorial revision)
- Steel MA, Huson D, Lockhart PJ (2000) Invariable site models and their use in phylogeny reconstruction. *Mol. Biol. Evol.* 17: (in press)
- 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, Cao Y, Hauf, J, Hasegawa M (1999) Using novel phylogenetic methods to evaluate mammalian mtDNA. *Syst. Biol.* 48:31-53
- Waddell PJ, Penny D, Moore T (1997) Hadamard conjugations and modeling sequence evolution with unequal rates across sites. *Mol Phylog Evol* 8:33-50
- Waddell PJ, Steel MA (1997) General time-reversible distances with unequal rates across sites: Mixing Γ and Inverse Gaussian distributions with invariant sites. *Mol Phylog Evol* 8:398-414
- 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
- Zuckermandl E (1976) Evolutionary processes and evolutionary noise at the molecular level. *J. Mol. Evol.* 7:269-311

Captions for figures

Figure 1.

Different rates at codon positions. It is usually observed that the average rate of change is highest for codon position three. However, under the standard Kimura-Dickerson model (see 4b) where any nucleotide change is either neutral or lethal, sites are either variable or constant ('on' or 'off' in the figure). Thus, on this model the 1st and 2nd sites evolve more slowly on average, but each variable site saturates at the same rate (see Griffiths 1997). This simple biochemical model predicts 1st and 2nd positions of a codon are basically no more reliable than 3rd codon positions. In contrast, the covarion model explains a slower rate at 1st and 2nd positions as sites being on or off in different parts of the tree, thus potentially increasing their reliability.

Figure 2.

Expected accuracy versus time of divergence. Data was generated under a Kimura 2-ST model for nine taxa and the tree inferred by neighbor-joining using Phylip. Each point on the figure 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 320 Mya the correct tree was not recovered.

Figure 3.

Rate of saturation when two or four nucleotides are viable at a site. The probability of a site supporting the correct tree are `correc_2` (for 2 nucleotides viable) and `correc_4` (for 4 nucleotides viable). Similarly, probabilities for sites supporting the wrong tree are `wrong_2` and `wrong_4`. Sites are expected to saturate at about the same rate in both cases. For 2-states viable there are only 8 patterns at a site compared with 64 (three of which support directly the correct tree) when all nucleotides are viable.

Figure 4.

A difference between the covarion model (A) and rates across sites models (B & C). Under a covarion model, a site can change from the standard rate (solid line) to zero (dashed line), and *vice versa*. In the rates across sites model (B), the rate at any site is the same over the whole tree, but there is a distribution of sites evolving at different rates (faster in 3B, slower in 3C).

Figure 5.

Parameters for a hidden Markov model for nucleotide evolution. (A) The instantaneous rate matrix \mathbf{K}' , and (B) a graphical representation. The diagonals (labeled *) are set so that each row of the rate matrix sums to 0. The arrows on the graphical form correspond to the positive entries in the rate matrix. The rates from A^+ (α , β , γ and δ) are shown on the graph.

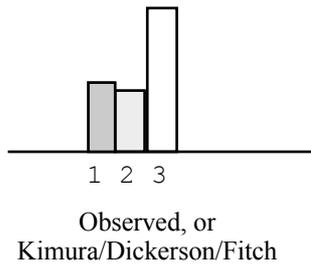
Figure 6.

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

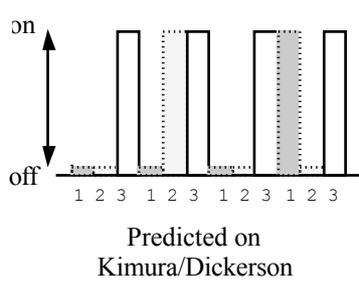
Figure 7.

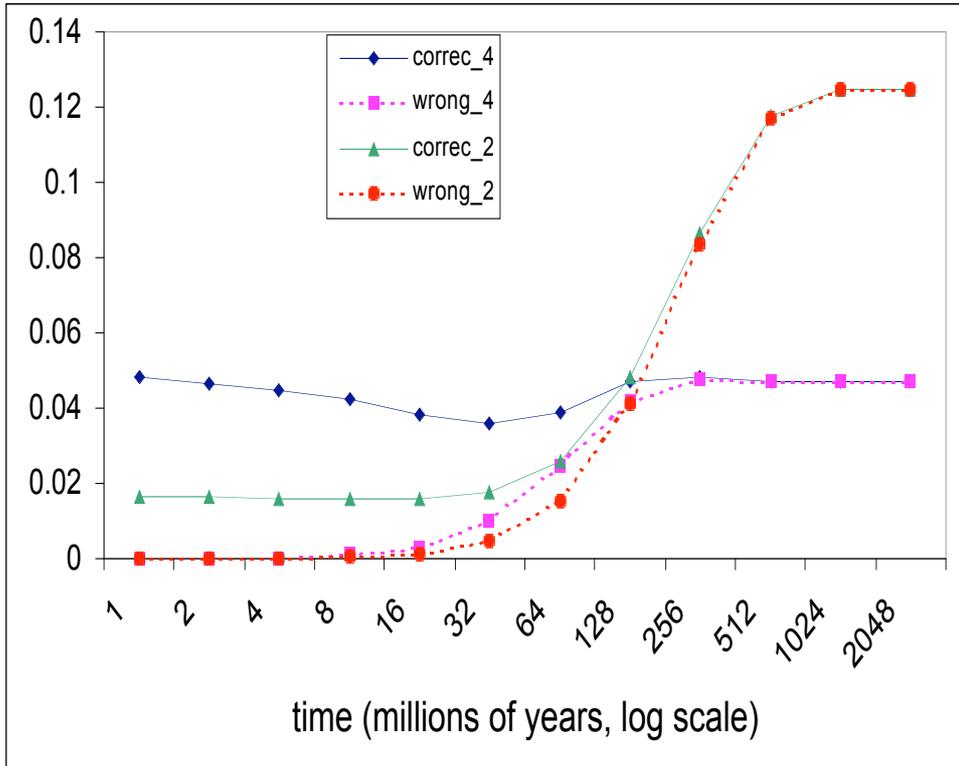
Results for neighbor-joining (A) and maximum likelihood (B) for data simulated under a covarion model. Each point represents 1000 simulations for sequences of 1000 nucleotides on the tree in Fig. 6B (see caption for Fig 6). The x-axis represents times since divergence (on a logarithmic scale) and the y-axis the probability of recovering the generating tree correctly. On each figure, the seven curves are δ values (the rate of interconversion between fixed and variable sites).

a



b





Covariation

Rates across sites

A

B 'fast'

C 'slow'

