THE COVARION MODEL OF MOLECULAR EVOLUTION

BENNET M^{c}COMISH

1997

THE COVARION MODEL OF MOLECULAR EVOLUTION

A thesis presented in partial fulfilment

of the requirements for the degree of

Master of Philosophy in Biology

at Massey University

Bennet M^{c}Comish

1997

Abstract

Current methods for constructing evolutionary trees generally do not work well for sequences in which multiple substitutions have occurred. The covarion hypothesis may provide a solution to this problem. This hypothesis states that only a limited number of the codons in a given sequence are free to vary, but that the set of variable codons may change as mutations are fixed in the population. Although this is reasonable from a biological point of view, it is a difficult hypothesis to test scientifically because the apparent large number of parameters involved makes it very hard to analyse statistically.

In this study, computer simulations were carried out on up to 51 machines running in parallel, using a simple covarion model based on a hidden Markov model (HMM) approach. This model required two new parameters—the proportion of sites that are variable at any given time, and the rate of exchange between fixed and variable states. These two parameters were both varied in the simulations. Sequence and distance data were simulated on a given tree under this covarion model, and these data were used to test the performance of standard tree-building methods at recovering the original tree

The neighbour joining and maximum likelihood methods tested were found to perform better with data generated under the covarion model than with data generated under a simpler model in which all sites vary at the same rate. This suggests that current tree-building methods may perform better with biological data than computer simulation studies suggest.

Acknowledgements

I am deeply grateful to everyone who has helped me, directly or indirectly, to get through this thesis.

I am most indebted to my supervisor, David Penny, without whose wise and patient guidance I would not have been able to complete this work. His ideas and programming skills were also crucial to the project.

I would also like to thank Mike Steel, who came up with the initial idea of using a hidden Markov model for the covarion hypothesis.

Various people here in the Department of Plant Biology and Something Else, and elsewhere, also deserve thanks for their ideas and other helpful things. These include Dan Jeffares, Anthony Poole, Kerryn Slack, Peter Waddell and Abby Harrison for numerous interesting discussions (in most cases completely unrelated to this project, but valuable none the less), Mike Hendy, Mike Steel, Pete Lockhart, Mike Charleston, Chris Tuffley and others for less frequent discussions of more immediate relevance, and Ted Drawneek for his assistance in setting up the framework for running simulations remotely and in parallel.

I thank all my friends (some of whom are mentioned above, but also Johan van Beek, Paul Hirst, Matt Barclay, and too many others to name here) for altering my sanity levels in whichever direction was necessary at the time, and not allowing me to stay in touch with reality for too much of the time.

I am eternally grateful to the various organisations to which I have sold my Eternal Soul™, and from which I have purchased Eternal Salvation™.

Finally, I thank both my parents for their support during this sometimes rather trying period of my life.

Table of Contents

List of Figures

List of Tables

List of Abbreviations and Symbols

2ST	Kimura two substitution-type model
3ST	Kimura three substitution-type model
covarion	Concomitantly variable codon
DNA	Deoxyribonucleic acid
HMM	Hidden Markov model
i.i.d.	Independent and identically distributed
SOD	Cu,Zn superoxide dismutase

Parameters used in the simulations:

α	Rate of transitions (A↔G and C↔T) in the 3ST model
β	Rate of type 1 transversions (A↔T and C↔G) in the 3ST model
γ	Rate of type 2 transversions (A↔C and G↔T) in the 3ST model
c	Sequence length
e	Rate of exchange between fixed and variable states
f	Proportion of sites that are permanently fixed
n	Number of taxa
r	Number of character states
s	Number of simulations to be performed
v	Proportion of sites that are variable

1. Introduction

1.1 Overview of the problem

A phylogenetic tree represents a hypothesis concerning the relationship between a set of taxa. Trees are constructed so that inferences can be made about the biology of the taxa. Tree-building methods can either use molecular sequence data directly, or they can use distances or dissimilarities between taxa.

Because events early in the history of life are not directly observable, it is difficult to know the accuracy of these evolutionary trees. Multiple substitutions at a site are an important source of error in the reconstruction of phylogenetic trees from sequence data, particularly when looking at older divergences. For example, if a sequence evolves at a rate of 3.3×10^{-9} substitutions per site per year (a realistic rate, Bulmer *et al.*, 1991), then after a billion years 5 changes would be expected to have occurred at each site in each lineage.

Some sites in the sequence may be conserved for functional reasons and do not change at all over time. Apart from assisting with alignment, these do not give information that can be used in selecting an optimal tree. At the other extreme, multiple substitutions can result in randomisation of the sequence at the sites that are free to vary, so that for ancient divergences these sites no longer contain information that can be used to infer relationships between the sequences. Computer simulations have shown that current methods are very unreliable when there has been an average of more than one change per site (Charleston, 1994, p.139).

Thus we appear to have a paradox. For sequences that diverged over a billion years ago, we expect sites to be either constant (and contain no relevant information), or so variable that they are saturated (and no longer contain any relevant information). The aim of this project is to investigate the covarion hypothesis as a means of finding a way out of this paradox.

In order to correct sequences for multiple substitutions we must assume a mechanism of sequence evolution. This mechanism will usually consist of a transition matrix determining the probability that nucleotide i at site x changes to nucleotide j along a given edge of the tree.

Many mechanisms have been suggested to describe sequence evolution, and these often assume that changes in the sequence are 'independent and identically distributed'(Penny *et al.*, 1992). That is, they assume all sites follow the same underlying process of substitution (identically distributed), and that a change at one site does not affect the chances of a change occurring at any other site, or in any other lineage (independent). These are known as i.i.d. models.

Commonly used models of sequence evolution include the Jukes-Cantor model (Jukes and Cantor, 1969) and the Kimura two and three substitution-type (2ST and 3ST) models (Kimura, 1980 and 1981 respectively). These models assume that changes are independent and identically distributed, and that the mechanism of change is constant over the whole tree. The Jukes-Cantor model assumes that all the rates of substitution from one base to another are equal, while Kimura's 2ST model allows for two rates, α and β, one each for transitions and transversions. The 3ST model is more general again, allowing two parameters, β and γ, for transversions (β for A$\leftrightarrow$T and G$\leftrightarrow$C; γ for A$\leftrightarrow$C and G$\leftrightarrow$T). Because all three models are symmetric (the rates of change in both directions between any two nucleotides, say A$\rightarrow$G and G$\rightarrow$A, are equal), the mean frequency of each of the four bases is expected to be the same and eventually end up as 1/4.

The 'identically distributed' assumption in most cases is not true, since it has been shown on many occasions that different sites in a macromolecule evolve at different rates (Fitch, 1971a; Uzzell and Corbin, 1971; Holmquist *et al.*, 1983; Tajima and Nei, 1984). There are two reasons for different rates of evolution at different sites—they may have either different mutation rates, or different selective constraints. The 'hot spots' observed in the D-loop of mitochondrial DNA (Wakeley, 1993) are an example of the former, but these

may be uncommon, at least in sequences used in phylogenetic analyses. Different selective constraints at different sites are likely to be much more common. In proteins, amino acid residues involved in the active site tend to evolve more slowly than other sites in the molecule, as predicted by Kimura's Neutral Theory of molecular evolution (Kimura, 1983), while for protein-coding DNA the third positions of codons evolve faster than the first and second.

Some models do take into account different rates for different codon positions, but still assume that changes are independent. However, a substitution at one site may alter the selective constraints on some other sites, allowing further substitutions, so in a formal sense changes are not independent. It has been shown that models that ignore correlated sites underestimate the actual amount of divergence (Schöniger and von Haeseler, 1994).

The covarion hypothesis, first proposed by Fitch and Markowitz in 1970, suggests a much more realistic, but also much more complex, mechanism. The complexity of this mechanism, however, makes it very difficult to analyse mathematically. This study uses a simplified covarion-style model as a basis for computer simulations in order to test the performance of tree-building methods on more realistic data than that produced by more widely used models.

1.2 The covarion hypothesis

Fitch and Markowitz (1970) observed that when 29 cytochrome *c* sequences from fungi, plants and animals were compared, 32 of the 113 codons were constant over all 29 taxa. When they reduced the range of species to non-primate mammals, however, the number of codons that were invariant rose to 95. They explained this by postulating 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. We shall use the term acceptable for such mutations. 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.*[*] "

This means that although most codons in a gene may be found to vary over a wide range of species, very few of the codons in a given species may be free to vary at a given point in time.

Fitch and Markowitz suggested that the interdependence of events at different coding positions may be related to the observations of Wyckoff (1968), who noted a spatial relationship between pairs of substitutions observed between rat and bovine ribonucleases.

In general, a mutation at one site in a protein is likely to alter the constraints on the molecule, so that some sites become free to change and others are no longer free to change. The sites affected in this way are likely to be close to the site in which the mutation occurs, that is, close in terms of the three-dimensional structure of the protein, not necessarily close in the sequence. This idea can be extended to the external constraints on the protein, so that a mutation in one protein may alter the covarion set of other proteins with which it interacts.

The covarion concept has also been extended to that of *co*ncomitantly *vari*able nucle*otides* or covariotides (Fitch, 1986), and applied to RNA, where the interactions between sites are mainly limited to complementary base pairing.

---

[*] The term 'covarions' was coined as an abbreviation of the phrase concomitantly variable codons.

1.3 History of the covarion hypothesis

In the two and a half decades since the covarion hypothesis was first proposed, only a handful of studies have analysed it in any detail. Notable examples of these detailed studies are Fitch (1971a), Karon (1979), Fitch and Ayala (1994), and Miyamoto and Fitch (1995).

Fitch (1971a) used a mathematical model for the covarion hypothesis to estimate the number of covarions, c, in cytochrome *c*, and the persistence of variability, v, of the covarions (where the persistence of variability is the probability of any given covarion retaining its variable status after a substitution elsewhere in the gene). This was done by examining the rate at which observable double mutations occur on a phylogenetic tree, and comparing this to expected values calculated for given values of c and v. The best fit was found to be between 4 and 10 covarions, with a persistence of variability of less than 0.25. This means that for each mutation fixed, 75% or more (on average) of the covarions lose their variable status.

Karon (1979) improved Fitch's mathematical 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. The same cytochrome *c* data was used as in Fitch and Markowitz (1970) and Fitch (1971a). This gave an average number of covarions of at most five, with about 35 to 65% of the covarions losing variability after each substitution. Karon also compared the covarion model with Holmquist, Cantor, and Jukes' random evolutionary hit (REH) interactive model (Holmquist *et al.*, 1972), and found that both models fit the data well, and thus both may be valid.

Fifteen years later, Fitch and Ayala (1994) used computer simulations to show that Cu,Zn superoxide dismutase (SOD), which had earlier been found to behave in an apparently very unclocklike manner (Ayala, 1986), could be a fairly accurate molecular clock under the covarion model, given an appropriate set of parameters. The parameters used were (*i*) sequence length of 118 potentially variable amino acids (out of a total of 162 codons), (*ii*) number of covarions = 28, (*iii*) persistence of variability = 0.01 (persistence of variability

has a different meaning in this paper than in the earlier studies, being the probability that no covarion will be exchanged for a presently invariable codon; only one of the covarions can be exchanged after each substitution—this makes it a somewhat more restrictive parameter), (*iv*) an average of 2.5 alternative amino acids at each variable site, and (*v*) 6 replacements per 10 million years.

Miyamoto and Fitch (1995) performed a more detailed simulation analysis of a subset of the SOD data of Ayala and Fitch, comparing the covarion model with both the Jukes-Cantor one-parameter model and the one-parameter process with a gamma distribution of rates across sites (Nei and Gojobori, 1986; Nei, 1991). The study focused on the difference between the varied and unvaried codons of mammals and plants, and found this to be more consistent with the covarion model than with either of the other models examined.

A few other papers have incorporated aspects of the covarion hypothesis into more general studies (for example Koonin and Gorbalenya, 1989; Fitch and Ye, 1991; Marshall *et al.*, 1994). Most papers that refer to covarions, however, only do so in passing, usually either pointing out possible examples of covarions (e.g. Bogardt *et al.*, 1976; Penny *et al.*, 1987), or simply citing the covarion hypothesis as established fact (e.g. Holmquist, 1972; Penny, 1974; Czelusniak *et al.*, 1978; Golding, 1983; Palumbi, 1989; Dorit and Ayala, 1995). A fairly comprehensive search of the literature only revealed one author (Gillespie, 1986 and 1988) who seems to disagree with the covarion hypothesis, out of over a hundred papers. Gillespie claims that the covarion model is simply an extreme example of a model in which some sites evolve more rapidly than others, apparently ignoring the main point of the covarion model, which is that different sites are free to change at different times.

1.4 The model

The simulations described in this thesis use a simplified covarion-style model. The covarion hypothesis as originally formulated was considered too complex because each site could be 'on' or 'off' (variable or fixed) on different parts of the tree. This appeared

to require a large number of parameters, since sites could only switch between the 'on' and 'off' states as a result of a substitution in one of the 'on' sites. With such a large number of parameters, almost any tree could be made to fit a given data set. One approach suggested as a way around the problem of too many parameters was to use a hidden Markov model (HMM).

Hidden Markov models (Baum and Petrie, 1966) have been applied to a number of problems in molecular biology over the last decade. Lander and Green (1987) used HMMs in the construction of genetic linkage maps, and they have also been used to distinguish coding from non-coding regions in DNA (Churchill, 1989). Simple HMMs have been used in conjunction with the Expectation-Maximisation algorithm to model certain protein-binding sites in DNA (Lawrence and Reilly, 1990; Cardon and Stormo, 1992). Protein families (Krogh *et al.*, 1994; Hughey and Krogh, 1996; Barrett *et al.*, 1997) and superfamilies (Stultz *et al.*, 1993) have been modelled using HMMs. HMMs have also been applied to multiple sequence alignment of proteins (Haussler *et al.*, 1993; Baldi *et al.*, 1994; Krogh *et al.*, 1994; Eddy,1995; Eddy *et al.*, 1995; Hughey and Krogh, 1996; McClure *et al.*, 1996; Barrett *et al.*, 1997), as well as protein structure prediction (Asai *et al,* 1993; Hubbard and Park, 1995). Mitchison and Durbin (1995) developed a tree-based HMM for maximum likelihood evolutionary trees which allows insertions and deletions, and Felsenstein and Churchill (1996) used an HMM to model variation in evolutionary rates among sites.

A hidden Markov model is a Markov model where the states of the system are not directly observable, but the observation is a probabilistic function of the state. That is, the HMM is a doubly embedded stochastic process with an underlying stochastic process that is hidden, but can be observed through another set of stochastic processes that produce the sequence of observations (Rabiner, 1989).

In the case of the covarion hypothesis, the 'hidden' part of the model would be changes between variable and fixed states at each site, and the observations would consist of the character-state (nucleotide or amino acid) at each site. Since this is such a simple example of an HMM, it can actually be reformulated as a single stochastic process.

The process modelled in these simulations (see figure 1) is based on the Kimura 3ST model, but with sites switching between fixed and variable states as well as between the four nucleotides. This gives us a total of eight character states, since each of the four nucleotides can be either fixed or variable at any given time. We will use the notations N_f and N_v (where N can be A, G, C, or T) for fixed and variable nucleotides respectively. This model only requires two additional parameters—the proportion of sites that are variable at any given time (that is, the number of covarions), and the rate of exchange between fixed and variable states. This model is essentially the same as that of Tuffley and Steel (1996), who analysed a simple covarion-style model based on an i.i.d. model but with an additional two-state Markov process that switches sites between the fixed and variable states. They showed that this model cannot be distinguished from one with a static distribution of rates across sites by pairwise comparison of sequences, but that the two models can be distinguished when there are at least four monophyletic groups of taxa.

This thesis explores the effects of varying the rate of exchange between fixed and variable states, as well as the proportion of sites that are covarions. Sequence and distance data are generated by computer simulation on a given tree, and these data are used to test the performance of two tree-building methods (maximum likelihood and neighbour joining) at recovering the original tree.

Even though these tree-building methods assume i.i.d. mechanisms, they are expected to be able to provide a good estimate of the correct tree after longer periods of time with data generated under the covarion-style model than with data generated under i.i.d. models, because more information will be lost due to multiple substitutions under an i.i.d. model. The neighbour joining and maximum likelihood methods tested were in fact found to perform better with data generated under the covarion model, suggesting that current tree-building methods may perform better with biological data than previous computer simulation studies have suggested.

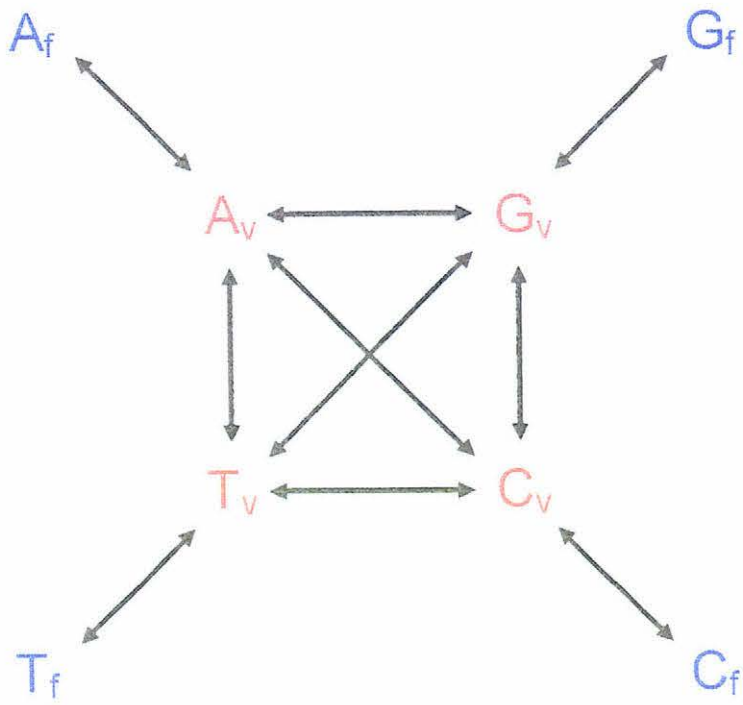

FIGURE 1: THE MODEL

The covarion-style model used in this study. The subscripts f and v denote fixed and variable states respectively.