12

Adaptive Molecular Evolution

Z. Yang

Department of Biology, Galton Laboratory, University College London, UK

This chapter reviews statistical methods for detecting adaptive molecular evolution by comparing synonymous and non-synonymous substitution rates in protein-coding DNA sequences. A Markov process model of codon substitution is introduced first, which forms the basis for all later discussion in this chapter. We then consider the case of comparing two sequences to estimate the numbers of synonymous (d_S) and non-synonymous (d_N) substitutions per site. The maximum likelihood (ML) method and a number of *ad hoc* methods are evaluated. The rest of the chapter deals with joint analyses of multiple sequences on a phylogeny. We review Markov models of codon substitution that allow the non-synonymous/synonymous rate ratio to vary among branches in a phylogeny or among amino acid sites in a protein. Those models can be used to construct likelihood ratio tests to identify evolutionary lineages under episodic Darwinian selection or to infer critical amino acid in a protein under diversifying selection. I use real data examples to demonstrate the application of the methods. The chapter finishes with a discussion of the limitations of current methods.

12.1 INTRODUCTION

While Darwin's theory of evolution by natural selection is accepted by biologists for morphological traits, the importance of selection in molecular evolution has been much debated. The neutral theory (Kimura, 1983) maintains that most observed molecular variation (both diversity within species and divergence between species) is due to random fixation of mutations with fitness effects so small that random drift rather than natural selection dominates their fate. Population geneticists have developed a number of tests of neutrality (see Wayne and Simonsen, 1998, for a review). Those tests often easily reject the strictly neutral model when applied to real data. However, they are often unable to distinguish different forms of natural selection, or to demonstrate molecular adaptation.

Up to now, the most convincing evidence of adaptive molecular evolution appears to have come from comparison of synonymous (silent) and non-synonymous (amino-acid-changing) substitution rates in protein-coding genes. We define the synonymous and

Handbook of Statistical Genetics. Edited by D.J. Balding et al.

non-synonymous rates (d_S and d_N) as the numbers of synonymous and non-synonymous substitutions per site, respectively. The ratio of the two rates, $\omega = d_N/d_S$, then measures selective pressure at the protein level. If selection has no effect on fitness, non-synonymous mutations will be fixed at the same rate as synonymous mutations, so that $d_N = d_S$ and $\omega = 1$. If non-synonymous mutations are deleterious, purifying selection will reduce their fixation rate, so that $d_N < d_S$ and $\omega < 1$. If non-synonymous mutations are favoured by Darwinian selection, they will be fixed at a higher rate that synonymous mutations, resulting in $d_N > d_S$ and $\omega > 1$. A significantly higher non-synonymous rate than the synonymous rate is thus evidence for adaptive evolution at the molecular level. This criterion has been used to identify several cases of positive selection, including the human major histocompatibility complex (Hughes and Nei, 1988), primate stomach lysozyme (Messier and Stewart, 1997), abalone sperm lysin (Lee et al., 1995), vertebrate visual pigments (Miyamoto and Miyamoto, 1996), and HIV-1 *env* genes (Bonhoeffer et al., 1995; Mindell, 1996; Yamaguchi and Gojobori, 1997). While still rare, those cases provide important insights into the mechanisms of molecular evolution.

The ω ratio has almost always been calculated as an average over all codons (amino acids) in the gene and over the entire evolutionary time that separates the sequences. The criterion that such an average ω is greater than one is a very stringent one for detecting positive selection (Sharp, 1997; Crandall et al., 1999). Biological considerations suggest that many amino acids in a protein are under strong functional constraints (with ω close to zero). Many proteins also appear to be under purifying selection during most of the evolutionary history. Adaptive evolution most likely occurs at a few time points and affects only a few amino acids (e.g. Stewart et al., 1987). In such a case, the ω ratio averaged over time and over sites will not be greater than 1 even if Darwinian selection has operated. For example, Endo et al. (1996) performed a large-scale database search and identified genes for which $d_N > d_S$ in at least half of the pairwise comparisons as potential targets for positive selection. Their analysis identified 17 proteins out of 3595, a proportion of only 0.47%. The scarcity of well-established cases of molecular adaptation appears partly due to the lack of power of the detection methods.

A remedy for this problem is to examine the ω ratio over a short evolutionary time period or in a short stretch of the gene such as functionally important domains. For example, Messier and Stewart (1997) used inferred ancestral genes to calculate the d_N and d_S rates for each branch in the phylogeny and identified two lineages that went through positive selection. Hughes and Nei (1988) found that the ω ratio is greater than 1 in a 57-codon region of the major histocompatibility complex that codes for the antigen-recognition site, although the ratio is less than 1 in other regions of the gene. Recently, likelihood models have been developed that account for variable ω ratios among branches in the phylogeny (Yang, 1998; Yang and Nielsen, 1998). Such models can be used to construct likelihood ratio tests of adaptive evolution along specific lineages, and have the advantage of not relying on inferred ancestral sequences. Models have also been developed that allow the ω ratio to vary among amino acid sites (Nielsen and Yang, 1998; Yang et al., 2000). Those models do not require knowledge of functionally important domains and may be used to test for the presence of critical amino acids under positive selection, and, when they exist, to identify them.

This chapter reviews statistical methods for phylogenetic analysis of protein-coding DNA sequences, with a focus on comparing synonymous and non-synonymous substitution rates to understand the mechanisms of sequence evolution. First, the probability

theory of the Markov process of codon substitution is briefly introduced. This theory forms the basis for maximum likelihood (ML) estimation of the d_N and d_S rates between two sequences as well as ML analysis of multiple sequences on a phylogeny. I will discuss different methods for comparing two sequences to estimate d_N and d_S. Besides ML (Goldman and Yang, 1994), there are about a dozen *ad hoc* methods for this estimation (e.g. Miyata and Yasunaga, 1980; Nei and Gojobori, 1986; Li et al., 1985; Li, 1993; Ina, 1995; Yang and Nielsen, 2000). These will be evaluated. The discussion will then turn to models that account for variable ω ratios among lineages and among sites, and use real data examples to explain their use in ML analysis. This chapter uses ML as the general framework. ML is known to have nice statistical properties, and indeed offers insights into *ad hoc* methods as well, which may not be based on an explicit probabilistic model. I will provide a brief introduction to ML estimation and the likelihood ratio test in Section 12.3. For a detailed and rigorous treatment, the reader may consult a statistics textbook such as Edwards (1992), Kalbfleisch (1985), or Stuart et al. (1999).

12.2 MARKOV MODEL OF CODON SUBSTITUTION

In molecular phylogenetics, we use a Markov process to describe the change between nucleotides, amino acids, or codons over evolutionary time. See Yang (1994), Swofford et al. (1996), and Liò and Goldman (1998) for use of Markov processes to model nucleotide substitution. In this chapter, our focus is the analysis of protein-coding DNA sequences, and the unit of evolution is a codon in the gene. We use a Markov process to describe substitutions between the sense codons. We exclude stop codons as they are usually not allowed in a protein. With the 'universal' genetic code, there are 61 sense codons (and 3 stop codons) and thus 61 states in the Markov process.

The Markov process is characterized by a rate (generator) matrix $Q = \{q_{ij}\}$, where q_{ij} is the substitution rate from sense codon i to sense codon $j (i \neq j)$. Formally, $q_{ij}\Delta t$ is the probability that the process is in state j after an infinitesimal time Δt, given that it is in state i at time t. The basic model we use in this chapter is simpler than the model of Goldman and Yang (1994) but more complex than that of Muse and Gaut (1994). It accounts for the transition/transversion bias, unequal synonymous and non-synonymous substitution rates, and biased base/codon frequencies. Mutations are assumed to occur independently among the three codon positions, and so only one position is allowed to change instantaneously. Since transitions (changes between T and C and between A and G) are known to occur more frequently than transversions (all other changes), we multiply the rate by the transition/transversion rate ratio, κ, if the change is a transition. Typical estimates of this parameter are 1.5–5 for nuclear genes and 5–30 for mitochondrial genes. To account for the codon usage bias, we let π_j be the equilibrium frequency of codon j and multiply substitution rates to codon j by π_j. We can either use all π_j as parameters, with 60 $(= 61 - 1)$ free parameters used, or calculate π_j from base frequencies at the three codon positions, with $9 = 3 \times (4 - 1)$ free parameters used.

To account for unequal synonymous and non-synonymous substitution rates, we multiply the rate by ω if the change is non-synonymous; ω is thus the non-synonymous/synonymous rate ratio, also termed the 'acceptance rate' by Miyata et al. (1979). In models we consider here, the relationship $\omega = d_N/d_S$ holds. For most genes, estimates of ω are much less than 1. It is important to note that parameters κ and π_j characterize processes at the DNA level

(which we call 'mutations', following Ina, 1995), while selection at the protein level has the sole effect of modifying parameter ω.

Formally, the substitution rate from codon i to codon j ($i \neq j$) is

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at two or three codon positions,} \\ \mu\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transversion,} \\ \mu\kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transition,} \\ \mu\omega\pi_j, & \text{if } i \text{ and } j \text{ differ by a non-synonymous transversion,} \\ \mu\omega\kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a non-synonymous transition.} \end{cases} \quad (12.1)$$

For example, consider substitution rates to codon CTG (which encodes amino acid Leu). We have $q_{\text{CTC, CTG}} = \mu\pi_{\text{CTG}}$ since the CTC (Leu) $\rightarrow$ CTG (Leu) change is a synonymous transversion, $q_{\text{TTG, CTG}} = \mu\kappa\pi_{\text{CTG}}$ since the TTG (Leu) $\rightarrow$ CTG (Leu) change is a synonymous transition, $q_{\text{GTG, CTG}} = \mu\omega\pi_{\text{CTG}}$ since the GTG (Val) $\rightarrow$ CTG (Leu) change is a non-synonymous transversion, and $q_{\text{CCG, CTG}} = \mu\kappa\omega\pi_{\text{CTG}}$ since the CCG (Pro) $\rightarrow$ CTG (Leu) change is a non-synonymous transition. Also $q_{\text{TTT, CTG}} = 0$ since codons TTT and CTG differ at two positions.

The diagonal elements of the rate matrix $Q = \{q_{ij}\}$ are determined by the mathematical requirement (e.g. Grimmett and Stirzaker, 1992, p. 241) that each row in the matrix sums to zero:

$$\sum_j q_{ij} = 0, \text{ for any } i. \quad (12.2)$$

Furthermore, molecular sequence data do not allow separate estimation of the rate (μ) and time (t), and only their product (μt) can be identified. We thus fix the rate μ such that the expected number of nucleotide substitutions per codon is one

$$-\sum_i \pi_i q_{ii} = \sum_i \pi_i \sum_{j \neq i} q_{ij} = 1. \quad (12.3)$$

This scaling means that time t is measured by distance, the expected number of (nucleotide) substitutions per codon. The transition probability matrix over time t is

$$P(t) = \{p_{ij}(t)\} = e^{Qt}, \quad (12.4)$$

where $p_{ij}(t)$ is the probability that codon i will become codon j after time t. As long as the rate matrix Q can be constructed, $P(t)$ can be calculated for any t using matrix diagonalization or Taylor expansion. Note that over any time interval, there is a non-zero probability that any codon i will change to any other codon j, even if they are separated by two or three nucleotide differences; that is, for any $t > 0$, $p_{ij}(t) > 0$ for any codons i and j.

Lastly, the model specified by equation (12.1) is time-reversible; that is, $\pi_i q_{ij} = \pi_j q_{ji}$ for any i and j. This means that

$$\pi_i p_{ij}(t) = \pi_j p_{ji}(t), \text{ for any } t, i \text{ and } j. \quad (12.5)$$

Note that $\pi_i p_{ij}(t)$ measures the amount of change from codons i to j over time t, while $\pi_j p_{ji}(t)$ measures the change in the opposite direction. Equation (12.5), known as the 'detailed balance', means that we expect to see equal numbers of changes from i to j and from j to i. I will mention some implications of reversibility in later sections.

12.3 ESTIMATION OF SYNONYMOUS AND NON-SYNONYMOUS SUBSTITUTION RATES BETWEEN TWO SEQUENCES

12.3.1 *Ad hoc* Methods

We want to estimate the number of synonymous substitutions per synonymous site (d_S) and the number of non-synonymous substitutions per non-synonymous site (d_N) between two protein-coding DNA sequences. In the past two decades, about a dozen *ad hoc* methods have been proposed for this estimation. Those methods are intuitive and involve treatment of the data that cannot be justified rigorously. Important basic concepts were developed in the early 1980s (Miyata and Yasunaga, 1980; Perler et al., 1980; Gojobori, 1983; Li et al., 1985; see also Ina, 1995, for a recent discussion), which we explain here with a hypothetical example. The critical question is by how much natural selection at the protein level has increased or decreased the non-synonymous substitution rate. Suppose the gene has 300 codons and we observe 5 synonymous and 5 non-synonymous differences (substitutions) between the two sequences. Can we conclude that synonymous and non-synonymous substitution rates are equal with $\omega = 1$? The answer is 'No'. An inspection of the genetic code table suggests that all changes at the second codon position and most changes at the first position are non-synonymous, and only some changes at the third position are synonymous. As a result, we do not expect to see equal proportions of synonymous and non-synonymous mutations even if there is no selection at the protein level. Indeed, if mutations from any one nucleotide to any other occur at the same rate, we expect 25.5% of mutations to be synonymous and 74.5% to be non-synonymous (Yang and Nielsen, 1998). If we use those proportions, it is clear that selection at the protein level has decreased the fixation rate of non-synonymous mutations by about three times, since $\omega = (5/5)/(74.5/25.5) = 0.34$. There are 900 nucleotide sites in the sequence, so the numbers of synonymous and non-synonymous sites are $S = 900 \times 25.5\% = 229.5$ and $N = 900 \times 74.5\% = 670.5$, respectively. We then have $d_S = 5/229.5 = 0.0218$ and $d_N = 5/670.5 = 0.0075$.

All *ad hoc* methods roughly follow the above intuitive procedure (for reviews, see Ina, 1996; Yang and Nielsen, 2000). They involve three steps. The first step is to count the numbers of synonymous (S) and non-synonymous (N) *sites* in the two sequences; that is, the number of nucleotide sites in the sequence is classified into the synonymous and non-synonymous categories, measuring mutational opportunities. This step is complicated by factors such as transition/transversion rate bias and base/codon frequency bias, both of which are ignored in our hypothetical example. The second step is to count the numbers of synonymous and non-synonymous *differences* between the two sequences; that is, the observed differences between the two sequences are classified into the synonymous and non-synonymous categories. This is straightforward if the two compared codons differ at one codon position only. When they differ at two or three codon positions, there exist four or six pathways from one codon to the other. The multiple pathways may involve different numbers of synonymous and non-synonymous differences and should ideally be weighted appropriately, although most *ad hoc* methods use equal weighting. The third step is to apply a correction for multiple substitutions at the same site since an observed difference may be the result of two or more substitutions. In our hypothetical example, we ignored the possibility of multiple hits and treated the observed differences as substitutions. All *ad hoc* methods have used multiple-hit correction formulas

based on nucleotide-substitution models, which assume that each nucleotide can change to one of three other nucleotides. When those formulas are applied to synonymous (or non-synonymous) sites only, this basic assumption of the Markov model is violated. Nevertheless, such corrections appear usable when the sequence divergence is low.

The method of Miyata and Yasunaga (1980) and its simplified version (Nei and Gojobori, 1986) are based on the nucleotide-substitution model of Jukes and Cantor (1969), and ignore the transition/transversion bias or base/codon frequency bias. As transitions are more likely to be synonymous at the third positions than transversions are, ignoring the transition/transversion rate bias leads to underestimation of the number of synonymous sites and overestimation of the number of non-synonymous sites. This effect is well known, and a number of attempts have been made to account for different transition and transversion rates in counting sites and differences (Li et al., 1985; Li, 1993; Pamilo and Bianchi, 1993; Comeron, 1995; Ina, 1995). The effect of biased base/codon frequencies has not been appreciated until recently (Moriyama and Powell, 1997). Yang and Nielsen (1998; 2000) found that extremely biased base/codon frequencies can have devastating effects on estimation of d_N and d_S, even outweighing the effect of the transition/transversion bias. Yang and Nielsen (2000) incorporated both the transition/transversion bias and the base/codon frequency bias in their *ad hoc* method.

12.3.2 Maximum Likelihood Estimation

Maximum likelihood is a powerful and flexible methodology for estimating parameters and testing hypotheses. Since the data are observed, we view the probability of observing the data as a function (the likelihood function) of the unknown parameters. The likelihood or log-likelihood function is our inference tool and contains all information about the parameters in the model. We estimate the unknown parameters by maximizing the likelihood function. Furthermore, the log-likelihood value under a model measures the fit of the model to data, and we compare two models by comparing their log-likelihood values. This is known as the likelihood ratio test. When two models are nested, twice the log-likelihood difference between the two models can be compared with the χ^2 distribution with the degree of freedom given by the difference in the number of parameters between the two models. The χ^2 approximation to the likelihood ratio statistic relies on large sample sizes (long sequences). How large the sample should be for the χ^2 approximation to be reliable depends on the specific model being tested as well as other factors such as sequence divergence. In a few cases of likelihood ratio tests applied to phylogenetics examined by computer simulation, the χ^2 approximation appears very good with as few as 100 or 200 nucleotides in the sequence. When the sequences are too short or when the two models are not nested, the correct distribution of the test statistic can be derived by Monte Carlo simulation (Goldman, 1993).

Below we describe the ML method for estimating d_N and d_S of Goldman and Yang (1994). The data are two aligned protein-coding DNA sequences. As a numerical example, we will use the human and mouse acetylcholine receptor α genes. The first 15 codons of the gene are as follows:

```
Human   GAG CCC TGG CCT CTC CTC CTG CTC TTT AGC CTT TGC TCA GCT GGC ...

Mouse   GAG CTC TCG ACT GTT CTC CTG CTG CTA GGC CTC TGC TCC GCT GGC ...
```

We assume that different codons in the sequence are evolving independently according to the same Markov process. As a result, data at different sites are independently and

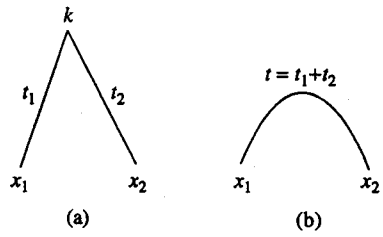

Figure 12.1 The tree for two sequences, with the codons x_1, x_2 for one site shown. Codon-substitution models considered in this chapter are all time-reversible and do not allow identification of the root. As a result, (a) parameters t_1 and t_2 cannot be estimated separately, and (b) only their sum $t = t_1 + t_2$ is estimable.

identically distributed. Suppose there are n sites (codons) in the gene, and let the data at site h be $\mathbf{x}_h = \{x_1, x_2\}$, where x_1 and x_2 are the two codons in the two sequences at that site (see Figure 12.1(a)). In the above example, the data at site $h = 2$ are $x_1 = \text{CCC}$ and $x_2 = \text{CTC}$. The probability of observing data $\mathbf{x}_h$ at site h is

$$f(\mathbf{x}_h) = \sum_{k=1}^{61} \pi_k p_{kx_1}(t_1) p_{kx_2}(t_2). \tag{12.6}$$

The term in the sum is the probability that the ancestor has codon k and the two current species have codons x_1 and x_2 at the site. This probability is equal to the prior probability that the ancestor has codon k, given by the equilibrium frequency π_k, multiplied by the two transition probabilities along the two branches of the tree (Figure 12.1(a)). Since the ancestral codon k is unknown, we sum over all possibilities for k. Time reversibility of the Markov process implies that

$$f(\mathbf{x}_h) = \sum_{k=1}^{61} \pi_{x_1} p_{x_1 k}(t_1) p_{kx_2}(t_2) = \pi_{x_1} \sum_{k=1}^{61} p_{x_1 k}(t_1) p_{kx_2}(t_2) = \pi_{x_1} p_{x_1 x_2}(t_1 + t_2). \tag{12.7}$$

The last step follows from the Chapman–Kolmogorov theorem (e.g. Grimmett and Stirzaker, 1992, pp. 239–246). Thus the statistical behaviour of the data is the same whether we consider the two sequences to be descendants of a common ancestor (as in Figure 12.1(a)) or we consider one sequence to be ancestral to the other (as in Figure 12.1(b)). In other words, the root of the tree cannot be identified, and only $t = t_1 + t_2$ can be estimated, but not t_1 and t_2 individually. Parameters in the model are the sequence divergence t, the transition/transversion rate ratio κ, the non-synonymous/synonymous rate ratio ω, and the codon frequencies π_j. The log-likelihood function is then given by

$$\ell(t, \kappa, \omega) = \sum_{h=1}^{n} \log\{f(\mathbf{x}_h)\}. \tag{12.8}$$

If some sites have the same data $\mathbf{x}$, the probability $f(\mathbf{x})$ needs to be calculated only once. An equivalent way of deriving the likelihood function is to note that the data follow a multinomial distribution with 61^2 categories corresponding to the 61^2 possible site patterns (configurations).

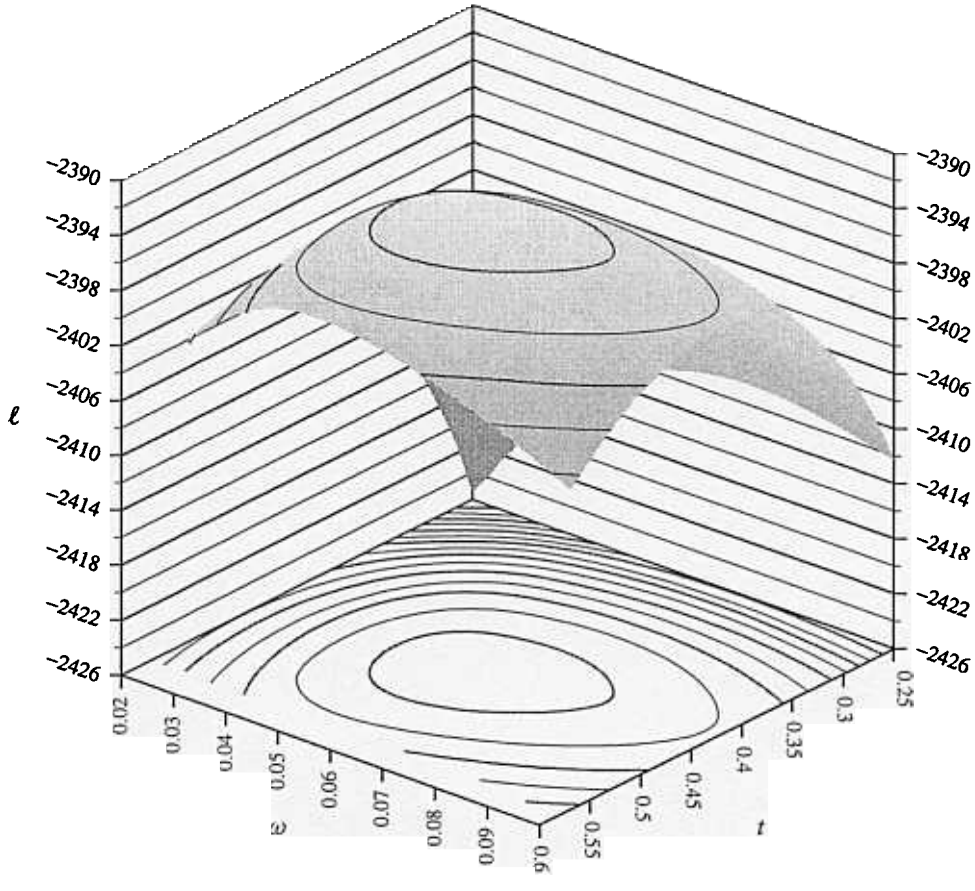

Figure 12.2 The log-likelihood surface contour as a function of parameters t and ω for the comparison of the human and mouse acetylcholine receptor α genes. The maximum likelihood method estimates parameters by maximizing the likelihood function. For these data, the estimates are $t = 0.444$, $\omega = 0.059$, with optimum log-likelihood $\ell = -2392.83$.

We usually estimate the codon frequencies (π_j) by the observed base/codon frequencies. To estimate parameters t, κ, and ω, we use a numerical hill-climbing algorithm to maximize ℓ, since an analytical solution is impossible. Figure 12.2 shows a log-likelihood surface as a function of t and ω for the human and mouse acetylcholine receptor α genes. The model assumes no transition/transversion bias or codon usage bias (with $\kappa = 1$ and $\pi_j = 1/61$ fixed), and involves two parameters only. This is the model underlying the method of Miyata and Yasunaga (1980) and Nei and Gojobori (1986).

The d_N and d_S rates are defined as functions of parameters t, κ, ω, and π_j, and their ML estimates are simply functions of ML estimates of parameters t, κ, ω, and π_j. The following description thus gives both the definitions of d_N and d_S and also the ML method for their estimation. The basic idea is the same as explained in our earlier hypothetical example. Here we count sites and substitutions per codon rather than for the entire sequence. First, note that the sequence divergence t is defined as the number of nucleotide substitutions per codon. We partition this number into the synonymous and

non-synonymous categories. We note that

$$\rho_S^* = \sum_{\substack{i \neq j \\ aa_i = aa_j}} \pi_i q_{ij} \tag{12.9}$$

and

$$\rho_N^* = \sum_{\substack{i \neq j \\ aa_i \neq aa_j}} \pi_i q_{ij} \tag{12.10}$$

are the proportions of synonymous and non-synonymous substitutions, respectively, as $\rho_S^* + \rho_N^* = 1$ (equation (12.3)). The summation in ρ_S^* is taken over all codon pairs i and j ($i \neq j$) that code for the same amino acid, while the summation in ρ_N^* is taken over all codon pairs i and j ($i \neq j$) that code for different amino acids; aa_i is the amino acid encoded by codon i. The numbers of synonymous and non-synonymous substitutions per codon are then $t\rho_S^*$ and $t\rho_N^*$, respectively.

Next, we calculate the proportions of synonymous and non-synonymous *sites*. Let these be ρ_S^1 and ρ_N^1. As noted before, these measure the mutational opportunities before the operation of selection at the protein level, that is, when $\omega = 1$ (Goldman and Yang, 1994; Ina, 1995). They are calculated similarly to equations (12.9) and (12.10), using the (mutational) transition/transversion rate ratio κ and codon frequencies (π_j), except that $\omega = 1$ is fixed. We assume there are three nucleotide sites in a codon (see Yang and Nielsen, 1998, for a discussion of the effect of mutations to stop codons). The numbers of synonymous and non-synonymous sites per codon are then $3\rho_S^1$ and $3\rho_N^1$, respectively. The numbers of synonymous and non-synonymous substitutions per site are then $d_S = t\rho_S^*/(3\rho_S^1)$ and $d_N = t\rho_N^*/(3\rho_N^1)$, respectively. Note that $\omega = d_N/d_S = (\rho_N^*/\rho_S^*)/(\rho_N^1/\rho_S^1)$, where the numerator is the ratio of the numbers of (observed) substitutions while the denominator is the ratio of the (expected) numbers of mutations when $\omega = 1$.

While the basic concepts discussed in the hypothetical example underlie both the ML and the *ad hoc* methods for estimating d_N and d_S (and their ratio ω), significant differences exist between the two classes of methods. In the ML method, the probability theory (that is, calculation of the transition probabilities by equation (12.4)) accomplishes several difficult tasks in one step: estimating mutational parameters such as κ, correcting for multiple hits, and weighting evolutionary pathways between codons. The Chapman–Kolmogorov theorem mentioned above states that $p_{ij}(t) = \sum_k p_{ik}(s)p_{kj}(t-s)$ for any $0 \leq s \leq t$; that is, the probability that codon i changes to codon j over time t is a sum over all possible codons (k) at any intermediate time point s. This theorem ensures that estimation of sequence divergence t from the likelihood function (equations (12.7) and (12.8)) accounts for all possible pathways of change between two codons, weighting them appropriately according to their relative probabilities of occurrence. When we partition the number of substitutions (t) into synonymous and non-synonymous categories, we only need to do it at the level of instantaneous rates (equations (12.9) and (12.10)), where there are no multiple changes.

In the *ad hoc* methods, each of the three steps offers a challenge. For example, some methods ignore the transition/transversion bias. Others take it into account but it has been difficult to estimate κ reliably. Ina (1995) used the third codon positions and Yang and Nielsen (2000) used so-called fourfold degenerate sites and non-degenerate sites to estimate κ, assuming that substitutions at those sites are either not affected or affected equally

by selection at the protein level. Both methods use nucleotide-based correction formulas to estimate κ, which seem problematic. Use of a limited class of sites also leads to large sampling errors in the estimates. The steps of counting differences, weighting pathways, and correcting for multiple hits are extremely complicated, when we want to incorporate major features of DNA sequence evolution such as the transition/transversion bias and the base/codon frequency bias (Yang and Nielsen, 2000). Notably, the synonymous and non-synonymous status of a site changes over time and also with the nucleotides at other positions of the codon (Muse, 1996). As a result, nucleotide-substitution models used in *ad hoc* methods are not capable of dealing with the complexity of the codon-substitution process.

12.3.3 A Numerical Example and Evaluation of Methods

To see the differences among methods for estimating d_N and d_S, we compare the human and mouse acetylcholine receptor α genes, using ML as well as several *ad hoc* methods (Table 12.1). The data set is the first (alphabetically) of the 49 genes analysed by Ohta (1995). The sequence has 456 codons (1368 nucleotides) after the start and stop codons are removed. With the ML method, we examine the effects of model assumptions. Some models ignore the transition/transversion rate ratio (with $\kappa = 1$ fixed) while others account for it (with κ estimated). Some ignore the codon frequency bias (Fequal) while others account for it to some extent (F1 × 4, F3 × 4, and F61) (see the note to Table 12.1 for definitions of these models).

Most of these models are nested, and the χ^2 approximation can be used to perform likelihood ratio tests. For example, we can compare models A and B in Table 12.1 to test whether there is transition/transversion rate bias. Model A is the null hypothesis and assumes that transition and transversion rates are equal ($\kappa = 1$). Model B does not

Table 12.1 Estimation of d_N and d_S between the human and mouse acetylcholine receptor α genes.

Model	κ	S	d_N	d_S	$d_N/d_S(\omega)$	ℓ
Ad hoc methods						
Nei and Gojobori (1986)		321.2	0.030	0.523	0.058	
Li (1993)	N/A	N/A	0.029	0.419	0.069	
Ina (1995)	6.1	408.4	0.033	0.405	0.081	
Yang and Nielsen (2000)	2.1	311.2	0.029	0.643	0.045	
ML methods						
(A) Fequal, $\kappa = 1$		348.5	0.029	0.496	0.059	−2392.83
(B) Fequal, κ estimated	2.8	396.7	0.031	0.421	0.073	−2379.60
(C) F1 × 4, $\kappa = 1$ fixed		361.0	0.029	0.513	0.057	−2390.35
(D) F1 × 4, κ estimated	2.9	406.5	0.031	0.436	0.071	−2376.12
(E) F3 × 4, $\kappa = 1$ fixed		281.4	0.029	0.650	0.044	−2317.72
(F) F3 × 4, κ estimated	3.0	328.1	0.030	0.545	0.055	−2303.33
(G) F61, $\kappa = 1$ fixed		261.5	0.028	0.736	0.038	−2251.92
(H) F61, κ estimated	3.0	319.5	0.030	0.613	0.048	−2239.33

Note: Fequal: equal codon frequencies (= 1/61) are assumed. F1 × 4: four nucleotide frequencies are used to calculate codon frequencies (3 free parameters). F3 × 4: nucleotide frequencies at three codon positions are used to calculate codon frequencies (9 free parameters). F61: all codon frequencies are used as free parameters (60 free parameters). ℓ is the log-likelihood value. Data are from Ohta (1995) and Yang and Nielsen (1998).

impose this constraint and has one more free parameter (κ) than model A. The likelihood ratio statistic, $2\Delta\ell = 2 \times (-2379.60 - (-2392.83)) = 2 \times 13.23 = 26.46$, should be compared with the χ^2 distribution with df = 1, giving a P-value of 0.27×10^{-6}. So there is significant difference between the transition and transversion rates.

For these data, both the transition/transversion bias and the codon frequency bias are clearly important. ML results under the most complex model (F61 with κ estimated), which accounts for both factors, are expected to be the most reliable and will be used to evaluate other methods/models. The F3 × 4 model is commonly used as it produces similar results to, and has far fewer parameters than, the F61 model. We note that *ad hoc* methods give similar results to ML under similar models; for example, Ina's method give similar estimates to ML accounting for the transition/transversion bias and ignoring the base/codon frequency bias (Table 12.1, method B).

It is well known that ignoring the transition/transversion rate bias leads to underestimation of the number of synonymous sites (S), overestimation of d_S, and underestimation of the ω ratio. This effect is obvious in Table 12.1 when ML estimates with κ estimated are compared with ML estimates when κ is fixed at 1, or when the method of Nei and Gojobori (1986) is compared with those of Li (1993) or Ina (1995). The effect of codon frequency bias is more complicated. In theory, biased base/codon frequencies can either increase or decrease the number of synonymous sites S (Yang and Nielsen, 2000). However, a scan over 80 mammalian nuclear genes, which include the 49 genes analysed by Ohta (1995), suggests that biased codon frequencies all lead to reduced numbers of synonymous sites. This is the pattern we see in Table 12.1, as estimates of S under the F3 × 4 and F61 models are much smaller than under the Fequal model. The gene is GC-rich at the third codon position, with base frequencies of 16% for T, 43% for C, 14% for A, and 27% for G. As a result, most mutations at the third codon position are transversions between C and G, and there are more non-synonymous mutations (sites) than expected under equal base/codon frequencies. In this data set, the effect of biased base frequencies is opposite to and outweighs the effect of the transition/transversion bias. As a result, the method of Nei and Gojobori (1986) *overestimates* rather than *underestimates* S and ω, contrary to general belief. The method of Ina (1995) accounts for the transition/transversion bias but ignores the codon frequency bias, and performs more poorly than the method of Nei and Gojobori (1986). The method of Yang and Nielsen (2000) accounts for both biases, and seems to produce estimates close to ML estimates under realistic models.

In general, different methods can produce either very similar or very different estimates of ω. With very weak transition/transversion bias and little codon usage bias, different methods tend to produce similar results. For other data sets, estimates from different methods can vary by a factor of 3–5 (Yang and Nielsen, 2000). Such large differences can occur even with highly similar sequences, as extreme transition/transversion bias or codon usage bias can drastically affect the counting of sites. One feature of the estimation is that when a method overestimates d_S, it tends to underestimate d_N at the same time, resulting in large errors in the ω ratio. This is because the total number of sites (or differences) is fixed, and if the method underestimates the number of synonymous sites (or differences) it will overestimate the number of non-synonymous sites as well, and vice versa.

Simulation studies performed to compare different methods produced results that are consistent with real data analysis. For example, Ina (1995) compared several *ad hoc* methods and concluded that none of them performed well when base frequencies are

extreme. Yang and Nielsen (2000) examined the effects of the transition/transversion rate bias and base/codon frequency bias, and found that estimation of the ω ratio is very sensitive to both biases. A worrying result is that the method of Nei and Gojobori (1986) can both underestimate and overestimate the ω ratio, often with large biases. In general, *ad hoc* methods may be used for exploratory data analysis, and the ML method accounting for both the transition bias and the codon usage bias should be preferred. It will be most sensible to use ML to analyse all sequences on a phylogeny simultaneously, if that is computationally feasible.

12.4 LIKELIHOOD CALCULATION ON A PHYLOGENY

Likelihood calculation for multiple sequences on a phylogeny may be viewed as an extension of the calculation for two sequences. The calculation is also similar to that under a nucleotide-substitution model (Felsenstein, 1981), although we now consider a codon rather than a nucleotide as the unit of evolution. We assume in this section that the same rate matrix Q (equation (12.1)) applies to all lineages and all amino acid sites. The data are multiple aligned sequences from different species. We assume independent substitutions among sites (codons), so that data at different codon sites are independently and identically distributed. The likelihood is given by the multinomial distribution with 61 categories (site patterns) for species. Let n be the number of sites (codons) in the sequence and the data at site h be $\mathbf{x}_h$ ($h = 1, 2, \ldots, n$); $\mathbf{x}_h$ is a vector of observed codons in different sequences at site h. An example tree of four species is shown in Figure 12.3. As in the case of two sequences, the root cannot be identified, and is arbitrarily fixed at the node ancestral to sequences 1 and 2. The data $\mathbf{x}_h$ can be generated by any codons j and k for the two ancestral nodes in the tree, and thus the probability of observing the data is a sum over all such possibilities:

$$f(\mathbf{x}_h) = \sum_j \sum_k [\pi_j p_{jx_1}(t_1) p_{jx_2}(t_2) p_{jk}(t_0) p_{kx_3}(t_3) p_{kx_4}(t_4)]. \qquad (12.11)$$

The quantity in the square bracket is the contribution to $f(\mathbf{x}_h)$ from ancestral codons j and k, and is equal to the prior probability that the codon at the root is j, which is given by the equilibrium frequency π_j, multiplied by the five transition probabilities along the five branches of the phylogeny (Figure 12.3). For a tree of s species with $s-2$ ancestral nodes, the data at each site will be a sum over 61^{s-2} possible combinations of ancestral codons. In computer programs, we use the pruning algorithm of Felsenstein (1981) to achieve efficient computation.

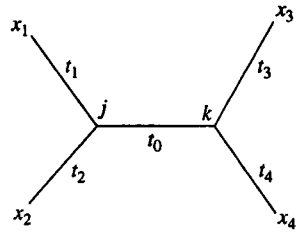

Figure 12.3 A tree of four sequences with codons at one site for nodes in the tree. Branch lengths $t_0, t_1, \ldots, t_4$ are parameters in the model.

The log-likelihood is a sum over all sites in the sequence:

$$\ell = \sum_{h=1}^{n} \log\{f(\mathbf{x}_h)\}. \tag{12.12}$$

Compared with the case of two sequences, we now have the same parameters in the substitution model (κ, ω, and the π_j), but many more branch length parameters (e.g. $t_0, t_1, \ldots, t_4$ in Figure 12.3 instead of the single t in Figure 12.1(b)). Again, numerical optimization algorithms have to be used to obtain ML estimates of parameters by maximizing the likelihood function.

As mentioned above, the quantity in the square bracket in equation (12.11) is the contribution to the probability of the data $f(\mathbf{x}_h)$ by ancestral codons j and k. This contribution varies greatly depending on the values of j and k, and the codons j and k that make the greatest contribution are the most probable codons for the two ancestral nodes at the site. This gives the empirical Bayes approach (also known as the likelihood approach) to reconstructing ancestral character states (Yang et al., 1995; Koshi and Goldstein, 1996). Compared with the parsimonious reconstruction (Fitch, 1971; Hartigan, 1973), the Bayes approach uses branch lengths and relative substitution rates between character states. *Ad hoc* methods that use reconstructed ancestral sequences to detect adaptive molecular evolution will be discussed later in comparison with the ML method.

12.5 DETECTING ADAPTIVE EVOLUTION ALONG LINEAGES

12.5.1 Likelihood Calculation Under Models of Variable ω Ratios Among Lineages

The major motivation for implementing models of variable ω ratios among lineages is that adaptive evolution probably happens in an episodic fashion. In a short time interval, non-synonymous mutations, driven by natural selection, may get fixed at a higher rate than synonymous mutations; as a result, the ω ratio for such an evolutionary lineage may be greater than 1. It is easy to modify the model of the previous section to allow for variable ω ratios among branches in a phylogeny. The likelihood calculation under such a model proceeds in a similar way, except that the transition probabilities for different branches need to be calculated from different rate matrices (Q) generated using different ωs. Suppose we want to fit a model in which the branch for species 1 of Figure 12.4 has a different ω ratio (ω_1), while all other branches have the same 'background' ratio ω_0. To indicate the dependence of p upon ω, let $p_{ij}(t;\omega)$ denote the transition probability

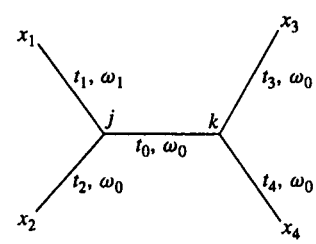

Figure 12.4 A tree of four sequences to explain a model of variable ω ratios among lineages. The ω ratio for the branch leading to species 1 (ω_1) is different from the ratio (ω_0) for all other branches.

calculated using the ratio ω. Under this model, the probability of observing data $\mathbf{x}_h$ is

$$f(\mathbf{x}_h) = \sum_j \sum_k \pi_j p_{jx_1}(t_1;\omega_1) p_{jx_2}(t_2;\omega_0) p_{jk}(t_0;\omega_0) p_{kx_3}(t_3;\omega_0) p_{kx_4}(t_4;\omega_0) \quad (12.13)$$

(compare with equation (12.11)). For the branch for species 1, ω_1 is used to construct the rate matrix Q and to calculate the transition probabilities, while for all other branches, ω_0 is used. Thus the model involves relatively minor modifications to the likelihood calculation under the model of one ω ratio for all branches in the tree, discussed above.

Yang (1998) implemented models that allow for different levels of heterogeneity in the ω ratio among lineages. The simplest model (the 'one-ratio' model) assumes the same ω ratio for all branches in the phylogeny. The most general model (the 'free-ratio' model) assumes an independent ω ratio for each branch in the phylogeny. Intermediate models such as two- or three-ratios models assume two or three different ω ratios for lineages in the tree. Those models can be compared using the likelihood ratio test to examine interesting hypotheses. For example, the likelihood values under the one-ratio and free-ratio models can be compared to test whether the ω ratios are different among lineages. Also we can allow the lineages of interest to have a different ω ratio from the background ω ratio for all other lineages in the phylogeny (as in Figure 12.4). Such a two-ratio model can be compared with the one-ratio model to examine whether the lineages of interest have a different ω ratio from other lineages. Furthermore, when the estimated ω ratio for the lineages of interest (say, ω_1 in Figure 12.4) is greater than 1, models with and without the constraint that $\omega_1 = 1$ can be compared to test whether the ratio is different from (i.e. greater than) 1. This test directly examines the possibility of positive selection along specific lineages.

It should be pointed out that variation in the ω ratio among lineages is a violation of the strictly neutral model, but is itself not sufficient evidence for adaptive evolution. First, relaxed selective constraints along certain lineages can generate variable ω ratios. Second, if non-synonymous mutations are slightly deleterious but not lethal, their fixation probabilities will depend on factors such as the population size of the species. In large populations, deleterious mutations will have a smaller chance of getting fixed than in small populations. Under such a model of slightly deleterious mutations (Ohta, 1973), species with large population sizes are expected to have smaller ω ratios than species with small population sizes. At any rate, an ω ratio significantly greater than 1 is unequivocal evidence of Darwinian selection.

12.5.2 Adaptive Evolution in the Primate Lysozyme

In the following, we use the example of the lysozyme *c* genes of primates (Figure 12.5) to demonstrate the use of codon substitution models of variable ω ratios among lineages (Yang, 1998). Lysozyme is found mainly in secretions such as tears and saliva as well as in white blood cells, where its function is to fight invading bacteria. Leaf-eating colobine monkeys have a complex foregut where bacteria ferment plant material, followed by a true stomach that expresses high levels of lysozyme, where its new function is to digest these bacteria (Stewart et al., 1987; Messier and Stewart, 1997). It has been suggested that the acquisition of a new function may have led to high selective pressure on the enzyme, resulting in high non-synonymous substitution rates. In an analysis of lysozyme *c* genes from 24 primate species, Messier and Stewart (1997) identified two lineages with

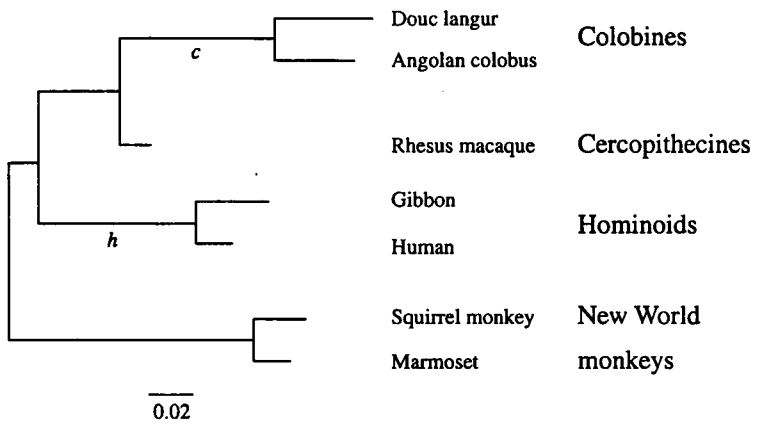

Figure 12.5 Phylogeny of seven primate species, for a subset of the lysozyme data set of Messier and Stewart (1997), used to demonstrate models of variable ω ratios among branches. After Yang (1998).

elevated ω ratios, indicating episodes of adaptive evolution in the lysozyme. One lineage, expected from previous analysis (Stewart et al., 1987), is ancestral to colobine monkeys, and another, unsuspected, lineage is ancestral to the hominoids. The two lineages are represented by branches h and c in Figure 12.5, for a subset of the data of Messier and Stewart (1997).

As branches h and c are the lineages of interest, we test assumptions concerning three ω ratio parameters: ω_h for branch h, ω_c for branch c, and ω_0 for all other (background) branches. Table 12.2 lists log-likelihood values and ML parameter estimates under different models. The simplest model (A) assumes one ω ratio, while the most general model assumes three ratios (E, I, and J). The three possible two-ratio models (B–D and F–H) are used as well. In models F–J, the ω ratio for the branch(es) of interest is fixed at 1.

Table 12.2 Log-likelihood values and parameter estimates under different models for the lysozyme c genes.

Model	p	ℓ	$\hat{\kappa}$	$\hat{\omega}_0$	$\hat{\omega}_h$	$\hat{\omega}_c$
A. 1-ratio: $\omega_0 = \omega_h = \omega_c$	22	−906.02	4.5	0.81	$= \hat{\omega}_0$	$= \hat{\omega}_0$
B. 2-ratios: $\omega_0 = \omega_h, \omega_c$	23	−904.64	4.6	0.69	$= \hat{\omega}_0$	3.51
C. 2-ratios: $\omega_0 = \omega_c, \omega_h$	23	−903.08	4.6	0.68	∞	$= \hat{\omega}_0$
D. 2-ratios: $\omega_0, \omega_h = \omega_c$	23	−901.63	4.6	0.54	7.26	$= \hat{\omega}_H$
E. 3-ratios: $\omega_0, \omega_h, \omega_c$	24	−901.10	4.6	0.54	∞	3.65
F. 2-ratios: $\omega_0 = \omega_h, \omega_c = 1$	22	−905.48	4.4	0.69	$= \hat{\omega}_0$	1
G. 2-ratios: $\omega_0 = \omega_c, \omega_h = 1$	22	−905.38	4.4	0.68	1	$= \hat{\omega}_0$
H. 2-ratios: $\omega_0, \omega_h = \omega_c = 1$	22	−904.36	4.3	0.54	1	1
I. 3-ratios: $\omega_0, \omega_h, \omega_c = 1$	23	−902.02	4.5	0.54	∞	1
J. 3-ratios: $\omega_0, \omega_h = 1, \omega_c$	23	−903.48	4.4	0.54	1	3.56

Note: p is the number of parameters. All models include the following 21 common parameters: 11 branch lengths in the tree (Figure 12.5), 9 parameters for base frequencies at codon positions used to calculate codon frequencies, and the transition/transversion rate ratio κ. *Source*: Yang (1998).

The estimate of the ω ratio under the one-ratio model ($\omega_0 = \omega_h = \omega_c$) is 0.81, indicating that, on average, purifying selection dominates the evolution of the lysozyme. Estimates of ω_c for branch c range from 3.4 to 3.6 when ω_c is allowed free to vary (models B, E, and J). Estimates of ω_h are always infinite when ω_h is assumed to be a free parameter (models C, E, and I), indicating the absence of synonymous substitutions along branch h. The estimate of the background ratio ω_0 is 0.54 when ω_h and ω_c are not constrained to be equal to ω_0 (models D, E, I, and J).

The results of likelihood ratio tests are shown in Table 12.3. Tests A–E examine whether the ω ratio for the branch(es) of interest is different from (that is, greater than) the background ratio, while tests A′–E′ examine whether the ratio is greater than 1. For example, test E compares models B and E of Table 12.2 and examines the null hypothesis that $\omega_h = \omega_0$, with ω_c free to vary in both models; ω_h is significantly higher than ω_0 in this comparison. Such tests suggest that ω_h is significantly greater than the background ratio ω_0 ($P < 1\%$; Table 12.3, D and E) and also significantly greater than one ($P < 5\%$; Table 12.3, D′ and E′). Similar tests suggest that ω_c is significantly greater than ω_0 ($P < 5\%$; Table 12.3, C), but not significantly greater than 1 (P ranges from 17% to 20%; Table 12.3, B′ and C′). More detailed analyses of the data set can be found in Yang (1998).

12.5.3 Comparison with Methods Based on Reconstructed Ancestral Sequences

Evolutionary biology has had a long tradition of reconstructing characters in extinct ancestral species and using them as observed data in all sorts of statistical as well as *ad hoc* analyses. The MacClade program (Maddison and Maddison, 1982) provides both a convenient tool for ancestral reconstruction using different variants of the parsimony algorithm and also an excellent review of the many uses and misuses of ancestral reconstruction. For molecular data, statistical methods (Yang et al., 1995; Koshi and Goldstein, 1996) can be used to obtain more reliable ancestral reconstructions, taking into account branch lengths and relative substitution rates between characters (nucleotides, amino acids, or

Table 12.3 Likelihood ratio statistics ($2\Delta\ell$) for testing hypotheses concerning lysozyme evolution.

Hypothesis tested	Assumption made	Models compared	$2\Delta\ell$
A. $(\omega_h = \omega_c) = \omega_0$	$\omega_h = \omega_c$	A & D	8.78**
B. $\omega_c = \omega_0$	$\omega_h = \omega_0$	A & B	2.76
C. $\omega_c = \omega_0$	ω_h free	C & E	3.96*
D. $\omega_h = \omega_0$	$\omega_c = \omega_0$	A & C	5.88*
E. $\omega_h = \omega_0$	ω_c free	B & E	7.08**
A′. $(\omega_h = \omega_c) \leq 1$	$\omega_h = \omega_c$	D & H	5.46*
B′. $\omega_c \leq 1$	$\omega_h = \omega_0$	B & F	1.68
C′. $\omega_c \leq 1$	ω_h free	E & I	1.84
D′. $\omega_h \leq 1$	$\omega_c = \omega_0$	C & G	4.60*
E′. $\omega_h \leq 1$	ω_c free	E & J	4.76*

*Significant at the 5% level ($\chi^2(1) = 3.84$).
**Significant at the 1% level ($\chi^2(1) = 6.63$).
Source: Yang (1998).

codons) (see the discussion below equation (12.12)). Overall, reconstructed molecular sequences appear much more reliable than reconstructed morphological characters (Yang et al., 1995; Cunningham et al., 1998).

Messier and Stewart (1997) reconstructed ancestral sequences and used them to perform pairwise comparisons to calculate the d_N and d_S rates along branches in the tree. Their analysis pinpoints two particular lineages in the primate phylogeny that may have gone through adaptive evolution. Crandall and Hillis (1997) took the same approach in an analysis of relaxed selective constraints in the rhodopsin genes of eyeless crayfishes living deep under the ground.

A major difference between the ML method discussed in this section and the approach of ancestral reconstruction is obviously that ML uses all possible ancestral characters (such as codons j and k for the two ancestral nodes in the tree of Figures 12.3 and 12.4), while the approach of ancestral reconstruction uses only the most likely codons and ignores the others. Ancestral sequences reconstructed by both parsimony and likelihood involve random errors and systematic biases. One kind of bias is obvious if we use reconstructed ancestral sequences to estimate branch lengths, as both parsimony and likelihood tend to minimize the amount of evolution to select the most likely ancestral characters. Biases involved in estimation of d_N and d_S using reconstructed ancestral sequences are not well characterized. As far as I know, none of the methods that use reconstructed ancestral characters have attempted to correct for biases in ancestral reconstruction. Furthermore, pairwise comparisons along branches of the phylogeny may not be as reliable as a simultaneous comparison of all sequences by ML.

It appears advisable that ancestral reconstruction be used for exploratory data analysis and that ML be preferred in general. When the likelihood ratio test suggests adaptive evolution along certain lineages, ancestral reconstruction may be very useful to pinpoint the responsible amino acid sites. Indeed, a most interesting use of ancestral reconstruction is to provide ancestral proteins to be synthesized in the laboratory to examine their biochemical and physiological properties. Such studies of 'palaeobiochemistry' were envisaged by Pauling and Zuckerkandl (1963) – for reviews, see Golding and Dean (1998) and Chang and Donoghue (2000).

12.6 INFERRING AMINO ACID SITES UNDER DIVERSIFYING SELECTION

12.6.1 Likelihood Calculation Under Models of Variable ω Ratios Among Sites

Up to now, we have assumed that all amino acid sites in a protein are under the same selective pressure, with the same underlying non-synonymous/synonymous rate ratio (ω). While the synonymous rate may be homogeneous among sites, non-synonymous rates are well known to be highly variable. Most proteins have highly conserved amino acid positions at which the underlying ω ratio is close to zero. The requirement that the ω ratio, averaged over all sites in the protein, is greater than 1 is thus a very stringent criterion for detecting adaptive evolution. It would be much more realistic if we allowed the ω ratio to vary among sites.

We can envisage two cases, which require different statistical modelling. In case 1, we may know the different structural and functional domains of the protein, and can use such

information to classify amino acid sites in the protein into several classes. The different site classes are assumed have different ω ratios, which are parameters to be estimated by ML. Suppose we have K site classes, with the corresponding ω ratios $\omega_1, \omega_2, \ldots, \omega_K$. The likelihood calculation under this model is rather similar to that under the model of one ω ratio for all sites (equations (12.11) and (12.12)), except that the right ω ratio will be used to calculate the transition probabilities for data at each site. For example, if site h is from site class k ($k = 1, 2, \ldots, K$) with ratio ω_k, then $f(\mathbf{x}_h)$ of equation (12.11) will be calculated using ω_k. The likelihood is again given by equation (12.12).

In case 2, we know or assume that there are several heterogeneous site classes with different ω ratios, but we do not know which class each amino acid site belongs to. Our discussion below focuses on this case. The standard practice is to use a statistical distribution to account for the variation of the ω ratio among sites (Nielsen and Yang, 1998). We assume that the synonymous rate is homogeneous among sites, and only the non-synonymous rates are variable. Branch length t is defined as the expected number of nucleotide substitutions per codon, averaged over sites. Suppose amino acid sites fall into K classes, with proportions $p_1, p_2, \ldots, p_K$ and ω ratios $\omega_1, \omega_2, \ldots, \omega_K$. The number of categories K is fixed beforehand, and the ps and ωs are either treated as parameters or as functions of parameters in the ω distribution. To calculate the likelihood, we want to calculate the probability of observing data at each site, say data $\mathbf{x}_h$ at site h. The conditional probability of the data given ω_κ, $f(\mathbf{x}_h|\omega_k)$, can be calculated as described earlier (equation (12.11)). Since we do not know which class site h belongs to, we sum over all site classes (that is, over the distribution of ω):

$$f(\mathbf{x}_h) = \sum_{k=1}^{K} p_k f(\mathbf{x}_h|\omega_k). \tag{12.14}$$

This is the same practice as summing over the unknown ancestral codons in equation (12.6). The log-likelihood is a sum over all n sites in the sequence:

$$\ell = \sum_{h=1}^{n} \log\{f(\mathbf{x}_h)\}. \tag{12.15}$$

Parameters in the model include branch lengths in the tree, κ, π_j, and parameters in the distribution of ω among sites. As before, we estimate the codon frequency parameters by the observed frequencies, and estimate the other parameters by numerical optimization of the likelihood.

After ML estimates of model parameters are obtained, we can use the empirical Bayes approach to infer the most likely site class (and thus the ω ratio) for any site. The marginal probability of the data $f(\mathbf{x})$ (equation (12.14)) is a sum of contributions from each site class k, and the site class that makes the greatest contribution is the most likely class for the site. That is, the posterior probability that a site with data $\mathbf{x}_h$ is from site class k (with rate ratio ω_k) is

$$f(\omega_k|\mathbf{x}_h) = \frac{p_k f(\mathbf{x}_h|\omega_k)}{f(\mathbf{x}_h)} = \frac{p_k f(\mathbf{x}_h|\omega_k)}{\sum_j p_j f(\mathbf{x}_h|\omega_j)}. \tag{12.16}$$

When the ω estimates for some site classes are greater than 1, this approach can be used to identify sites from such classes, which are potential targets of diversifying selection. The posterior probability provides a measure of accuracy.

Nielsen and Yang (1998) implemented a few simple models that allow for variable ω ratios among sites. The 'neutral' model assumes two classes of sites: the conserved sites at which non-synonymous mutations cannot survive, so that $\omega_1 = 0$; and the completely neutral sites at which $\omega_2 = 1$. This model involves only one free parameter (p_1) in the ω distribution, since $p_2 = 1 - p_1$. A 'selection' model adds a third class of sites with the underlying ω ratio estimated from the data. This model involves three free parameters in the ω distribution: p_1, p_2, and ω_3. These models appear too simple to capture the complexity of the substitution process of various proteins. Yang et al. (2000) implemented several new models, including a general discrete model, which treats all the p_i and ω_i, $i = 1, \ldots, K$, as parameters subject to the only constraint that the p_i sum to 1. Those authors also implemented several continuous mixture distributions, using discrete approximations to facilitate the computation (Yang et al., 2000). I will use a few such models in the numerical example in the next subsection, with details given there. Extensive analysis of ten data sets by Yang et al. (2000) suggests that quite a few of the models implemented can be used to test for the presence of positively selected sites and to identify them when they exist.

12.6.2 Positive Selection in the HIV-1 *vif* Genes

An example data set of HIV-1 *vif* genes from 29 subtype-B isolates is used here to demonstrate the likelihood models of variable ω ratios among sites (Yang et al., 2000). The sequence has 192 codons. Several models are used in ML estimation, with the results shown in Table 12.4. I list only those parameters involved in the ω distribution, as other parameters (branch lengths in the phylogeny, the transition/transversion rate ratio κ, and the base frequencies at the three codon positions) are common to all models. The model codes are those used in the PAML program package (Yang, 1999).

Table 12.4 Likelihood values and parameter estimates under models of variable ω ratios among sites for HIV-1 *vif* genes.

Model code	ℓ	κ	d_N/d_S	Estimates of parameters
M0. one-ratio (1)	−3499.60	3.72	0.644	$\omega = 0.644$
M1. neutral (1)	−3413.07	3.78	0.575	$p_1 = 0.425$, $(\omega_1 = 0)$ $(p_2 = 0.575)$, $(\omega_2 = 1)$
M2. selection (3)	−3377.94	4.22	0.870	$p_1 = 0.404$, $(\omega_1 = 0)$ $p_2 = 0.511$, $(\omega_2 = 1)$ $(p_3 = 0.085)$, $\omega_3 = 4.220$
M3. discrete (5)	−3367.16	4.13	0.742	$p_1 = 0.604$, $\omega_1 = 0.108$ $p_2 = 0.325$, $\omega_2 = 1.211$ $(p_3 = 0.070)$, $\omega_3 = 4.024$
M7. beta (2)	−3400.45	3.55	0.440	$p = 0.176$, $q = 0.223$
M8. beta & ω (4)	−3370.66	4.02	0.687	$p_1 = 0.909$, $p = 0.222$, $q = 0.312$ $(p_2 = 0.091)$, $\omega = 3.385$

Note: The number of parameters in the ω distribution is given in parentheses after the model code. d_N/d_S is the average ω ratio over all sites in the gene. Parameters in parentheses are given to ease interpretation but they are not free parameters. *Source*: Yang et al. (2000).

The model of one ω ratio for all sites (M0) gives an average ω ratio of 0.644, indicating that on average purifying selection is the dominating force during the evolution of the gene. The selection model (M2) suggests about $p_3 = 8.5\%$ of sites are under strong positive selection with $\omega_3 = 4.2$. The discrete model (M3) suggests about $p_3 = 7\%$ of sites are under strong positive selection with $\omega_3 = 4.0$, while a large proportion ($p_2 = 33\%$) of sites are under weak positive selection or are nearly neutral with $\omega_2 = 1.2$. Both models have significantly higher likelihood values than models M0 (one-ratio) and M1 (neutral), indicating presence of sites under diversifying selection. Model M7 assumes that ω is drawn from a beta distribution $B(p, q)$. The beta distribution can take a variety of shapes (such as L-, inverted L-, U-, and inverted U-shapes) but is restricted to the interval (0, 1). It provides a useful null model to test for positive selection. The estimated $B(0.176, 0.223)$ distribution has a U shape, possibly because the ω ratios at some sites are greater than 1, which the beta distribution cannot accommodate. Model M8 adds an extra class of sites to the beta, with a free ω ratio estimated. So a proportion p_1 of sites have the ω ratio drawn from the beta distribution, while the remaining sites have the same unknown ratio ω. Estimates under this model (Table 12.3) suggest that 90.9% of sites are from the beta distribution $B(0.222, 0.312)$, while the remaining sites (9.1%) are under positive selection with $\omega = 3.4$. The likelihood ratio statistic for comparing M7 and M8 is $2\Delta\ell = 2 \times 29.79 = 59.58$, much greater than $\chi^2_{1\%}(2) = 9.21$. M8 thus fits the data significantly better than M7, and we conclude that the data contain a class of sites under positive selection.

Figure 12.6 plots the posterior probabilities for site classes at each site under the discrete model (M3). Parameter estimates under this model suggest that the ω ratios for the three site classes are 0.108, 1.211, and 4.024. The site classes are in the proportions 60.4%, 32.5%, and 7.0% (Table 12.4). Those are the prior probabilities for site classes for each site. The observed data will alter those probabilities considerably, so that the posterior probabilities are very different from the prior. For example, the posterior probabilities for site 1 are 0.993, 0.007, and 0.000, and site 1 is almost certainly a highly conserved site. In contrast, the probabilities at site 31 are 0.000, 0.030, and 0.970, and this site is most likely under strong diversifying selection (Figure 12.6).

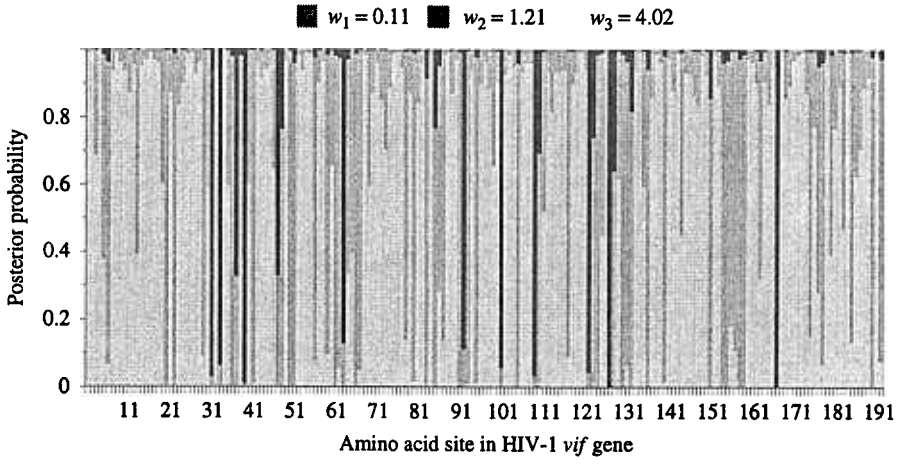

Figure 12.6 Posterior probabilities of site classes along the gene for the HIV-1 *vif* genes under the discrete model. After Yang et al. (2000).

12.6.3 Comparison with Methods Based on Reconstructed Ancestral Sequences

Reconstructed ancestral sequences can also be used to perform *ad hoc* analysis to identify amino acid sites in the protein that may be targets of diversifying selection. On a large phylogeny, it is possible to use reconstructed ancestral sequences to count the numbers of synonymous and non-synonymous changes along branches of the tree and compare them with 'neutral' expectations when selection at the protein level has no effect on the non-synonymous rate. However, the analysis suffers from errors and biases in inferred ancestral sequences, and it is problematic to devise rigorous statistical tests on data at one site. Fitch et al. (1997) performed such an intuitive analysis of the HA gene of human influenza virus type A. Suzuki and Gojobori (1999) developed a systematic approach based on this idea and were able to detect positively selected amino acid sites in a number of genes suspected to be undergoing adaptive evolution. Both studies used the parsimony algorithm to infer ancestral sequences and used them to count changes along branches in the tree for each site in the sequence. Such methods require a large number of sequences to operate, and may be useful for exploratory analysis of large data sets.

12.7 LIMITATIONS OF CURRENT METHODS

Both the test of positive selection along lineages and the test of positive selection at amino acid sites discussed in this chapter are highly conservative. When testing for lineages under positive selection, we assumed the ω ratio is identical across sites. Positive selection is detected along a lineage only if the ω ratio averaged over all sites is significantly greater than 1. Since many or most sites in a protein are under purifying selection with the underlying ω ratios close to 0, this procedure constitutes a very conservative test of positive selection. Similarly, the likelihood ratio test of positively selected sites is based on the assumption that the ω ratio is identical among all lineages on the tree. Positive selection is detected for a site only if the underlying ω ratio averaged over all lineages is significantly greater than 1. This assumption appears unrealistic for most genes, as positive selection probably affects only a few lineages.

The models discussed here assume the same ω ratio for any amino acid changes; at a positively selected site, changes to any amino acids are assumed to be advantageous. This assumption appears unrealistic for any protein. Furthermore, the tests discussed here only identify diversifying selection which increases the non-synonymous rates, and may have little power in detecting other types of selection such as balancing selection (Yang et al., 1999). While all those unrealistic assumptions make the test conservative for detecting adaptive evolution, it is noteworthy that in both the lysozyme gene and the HIV-1 *vif* gene examples, adaptive evolution is detected even though the average ω ratio is less than 1.

An obvious step to improve the power of the likelihood ratio tests is to develop models that allow the ω ratio to vary both among lineages and among sites. It appears possible to construct such a model, although it will require more computation than the models discussed here. Nevertheless, at some stage we will have to compromise. On the one hand, we want to focus on a short time period and a few amino acid sites so that the signal of adaptive evolution will not be overwhelmed by the effect of purifying selection during other time periods and at other amino acid sites. On the other hand, a short time

period and a few amino acid sites may not offer enough room for evolutionary change (adaptive or otherwise) to generate a signal of positive selection that is detectable by statistical tests.

12.8 COMPUTER SOFTWARE

Several programs are available to implement *ad hoc* methods to estimate d_N and d_S between two sequences; they are often distributed by the authors of the methods. PAML (Yang, 1999) is currently the only program that implements the ML models discussed in this chapter.

Acknowledgments

I thank David Balding, Joanna Holbrook, and Jennifer Wernegreen for comments. This work is supported by BBSRC grant 31/G10434.

REFERENCES

Bonhoeffer, S., Holmes, E.C. and Nowak, M.A. (1995). Causes of HIV diversity. *Nature* **376**, 125.

Chang, B.S. and Donoghue, M.J. (2000). Recreating ancestral proteins. *Trends in Ecology and Evolution* **15**, 109–114.

Comeron, J.M. (1995). A method for estimating the numbers of synonymous and non-synonymous substitutions per site. *Journal of Molecular Evolution* **41**, 1152–1159.

Crandall, K.A. and Hillis, D.M. (1997). Rhodopsin evolution in the dark. *Nature* **387**, 667–668.

Crandall, K.A., Kelsey, C.R., Imamichi, H., Lane, H.C. and Salzman, N.P. (1999). Parallel evolution of drug resistance in HIV: failure of non-synonymous/synonymous substitution rate ratio to detect selection. *Molecular Biology and Evolution* **16**, 372–382.

Cunningham, C.W., Omland, K.E. and Oakley, T.H. (1998). Reconstructing ancestral character states: a critical reappraisal. *Trends in Ecology and Evolution* **13**, 361–366.

Edwards, A.W.F. (1992). *Likelihood*, expanded edition. London: Johns Hopkins University Press.

Endo, T., Ikeo, K. and Gojobori, T. (1996). Large-scale search for genes on which positive selection may operate. *Molecular Biology and Evolution* **13**, 685–690.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17**, 368–376.

Fitch, W.M. (1971). Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Zoology* **20**, 406–416.

Fitch, W.M., Bush, R.M., Bender, C.A. and Cox, N.J. (1997). Long term trends in the evolution of H(3) HA1 human influenza type A. *Proceedings of the National Academy of Sciences (USA)* **94**, 7712–7718.

Gojobori, T. (1983). Codon substitution in evolution and the 'saturation' of synonymous changes. *Genetics* **105**, 1011–1027.

Golding, G.B. and Dean, A.M. (1998). The structural basis of molecular adaptation. *Molecular Biology and Evolution* **15**, 355–369.

Goldman, N. (1993). Statistical tests of models of DNA substitution. *Journal of Molecular Evolution* **36**, 182–198.

Goldman, N. and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution* **11**, 725–736.

Grimmett, G.R. and Stirzaker, D.R. (1992). *Probability and Random Processes*, 2nd edition. Clarendon Press, Oxford.

Hartigan, J.A. (1973). Minimum evolution fits to a given tree. *Biometrics* **29**, 53–65.

Hughes, A.L. and Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**, 167–170.

Ina, Y. (1995). New methods for estimating the numbers of synonymous and non-synonymous substitutions. *Journal of Molecular Evolution* **40**, 190–226.

Ina, Y. (1996). Pattern of synonymous and non-synonymous substitutions: an indicator of mechanisms of molecular evolution. *Journal of Genetics* **75**, 91–115.

Jukes, T.H. and Cantor, C.R. (1969). In *Mammalian Protein Metabolism*, H.N. Munro, ed. Academic Press, New York, pp. 21–123.

Kalbfleisch, J.G. (1985). *Probability and Statistical Inference, Vol. 2: Statistical Inference*. Springer-Verlag, New York.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

Koshi, J.M. and Goldstein, R.A. (1996). Probabilistic reconstruction of ancestral protein sequences. *Journal of Molecular Evolution* **42**, 313–320.

Lee, Y.H., Ota, T. and Vacquier, V.D. (1995). Positive selection is a general phenomenon in the evolution of abalone sperm lysin. *Molecular Biology and Evolution* **12**, 231–238.

Li, W.-H. (1993). Unbiased estimation of the rates of synonymous and non-synonymous substitution. *Journal of Molecular Evolution* **36**, 96–99.

Li, W.-H., Wu, C.-I. and Luo, C.-C. (1985). A new method for estimating synonymous and non-synonymous rates of nucleotide substitutions considering the relative likelihood of nucleotide and codon changes. *Molecular Biology and Evolution* **2**, 150–174.

Liò, P. and Goldman, N. (1998). Models of molecular evolution and phylogeny. *Genome Research* **8**, 1233–1244.

Maddison, W.P. and Maddison, D.R. (1982). *MacClade: Analysis of Phylogeny and Character Evolution*, 3rd edition. Sinauer, Sunderland, MA.

Messier, W. and Stewart, C.-B. (1997). Episodic adaptive evolution of primate lysozymes. *Nature* **385**, 151–154.

Mindell, D.P. (1996). Positive selection and rates of evolution in immunodeficiency viruses from humans and himpanzees. *Proceedings of National Academy of Sciences (USA)* **93**, 3284–3288.

Miyamoto, S. and Miyamoto, R. (1996). Adaptive evolution of photoreceptors and visual pigments in vertebrates. *Annual Review of Ecology and Systematics* **27**, 543–567.

Miyata, T. and Yasunaga, T. (1980). Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its applications. *Journal of Molecular Evolution* **16**, 23–36.

Miyata, T., Miyazawa, S. and Yasunaga, T. (1979). Two types of amino acid substitutions in protein evolution. *Journal of Molecular Evolution* **12**, 219–236.

Moriyama, E.N. and Powell, J.R. (1997). Synonymous substitution rates in Drosophila: mitochondrial versus nuclear genes. *Journal of Molecular Evolution* **45**, 378–391.

Muse, S.V. (1996). Estimating synonymous and non-synonymous substitution rates. *Molecular Biology and Evolution* **13**, 105–114.

Muse, S.V. and Gaut, B.S. (1994). A likelihood approach for comparing synonymous and non-synonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution* **11**, 715–724.

Nei, M. and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and non-synonymous nucleotide substitutions. *Molecular Biology and Evolution* **3**, 418–426.

Nielsen, R. and Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**, 929–936.

Ohta, T. (1973). Slightly deleterious mutant substitutions in evolution. *Nature* **246**, 96–98.

Ohta, T. (1995). Synonymous and non-synonymous substitutions in mammalian genes and the nearly neutral theory. *Journal of Molecular Evolution* **40**, 56–63.

Pamilo, P. and Bianchi, N.O. (1993). Evolution of the *Zfx* and *Zfy* genes – rates and interdependence between the genes. *Molecular Biology and Evolution* **10**, 271–281.

Pauling, L. and Zuckerkandl, E. (1963). Chemical paleogenetics: molecular 'restoration studies' of extinct forms of life. *Acta Chemica Scandinavica* 17, S9–S16.

Perler, F., Efstratiadis, A., Lomedica, P., Gilbert, W., Kolodner, R. and Dodgson, J. (1980). The evolution of genes: the chicken preproinsulin gene. *Cell* **20**, 555–566.

Sharp, P.M. (1997). In search of molecular Darwinism. *Nature* **385**, 111–112.

Stewart, C.-B., Schilling, J.W. and Wilson, A.C. (1987). Adaptive evolution in the stomach lysozymes of foregut fermenters. *Nature* **330**, 401–404.

Stuart, A., Ord, K. and Arnold, S. (1999). *Kendall's Advanced Theory of Statistics*, 6th edition, Vol. 2a. Arnold, London.

Suzuki, Y. and Gojobori, T. (1999). A method for detecting positive selection at single amino acid sites. *Molecular Biology and Evolution* **16**, 1315–1328.

Swofford, D.L., Olsen, G.J., Waddell, P.J. and Hillis, D.M. (1996). In *Molecular Systematics*, D.M. Hillis, C. Moritz and B.K. Mable, eds. Sinauer, Sunderland, MA, pp. 411–501.

Wayne, M.L. and Simonsen, K.L. (1998). Statistical tests of neutrality in the age of weak selection. *Trends in Ecology and Evolution* **13**, 236–240.

Yamaguchi, Y. and Gojobori, T. (1997). Evolutionary mechanisms and population dynamics of the third variable envelope region of HIV within single hosts. *Proceedings of the National Academy of Sciences (USA)* **94**, 1264–1269.

Yang, Z. (1994). Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution* **39**, 105–111.

Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular Biology and Evolution* **15**, 568–573.

Yang, Z. (1999). *Phylogenetic analysis by maximum likelihood (PAML)*. http://abacus.gene.ucl.ac.uk/software/paml.html.

Yang, Z. and Nielsen, R. (1998). Synonymous and non-synonymous rate variation in nuclear genes of mammals. *Journal of Molecular Evolution* **46**, 409–418.

Yang, Z. and Nielsen, R. (2000). Estimating synonymous and non-synonymous substitution rates under realistic evolutionary models. *Molecular Biology and Evolution* **17**, 32–43.

Yang, Z., Kumar, S. and Nei, M. (1995). A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* **141**, 1641–1650.

Yang, Z., Nielsen, R., Goldman, N. and Petersen, A.-M.K. (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**, 431–449.