REVIEW

Models of Molecular Evolution and Phylogeny

Pietro Liò and Nick Goldman[1]

Department of Genetics, University of Cambridge, Cambridge CB2 3EH, UK

Phylogenetic reconstruction is a fast-growing field that is enriched by different statistical approaches and by findings and applications in a broad range of biological areas. Fundamental to these are the mathematical models used to describe the patterns of DNA base substitution and amino acid replacement. These may become some of the basic models for comparative genome research. We discuss these models, including the analysis of observed DNA base and amino acid mutation patterns, the concept of site heterogeneity, and the incorporation of structural biology data, all of which have become particularly important in recent years. We also describe the use of such models in phylogenetic reconstruction and statistical methods for the comparison of different models.

PCR has deeply transformed and boosted phylogenetic studies. At the same time, the statistical analysis of evolutionary relationships among species has recently revealed important biotechnological uses. For example, the understanding of viral quasispecies variation allows us to trace routes of infectious disease transmission. The analysis of the host–pathogen relationships in terms of mutual genetic variation can led to a deeper insight into drug design for medical and agricultural purposes, and structural biologists are becoming interested in the phylogeny of sets of homologous proteins belonging to different organisms because these reflect all the different variants already experienced in nature and can reveal structural and functional constraints.

How does a geneticist reconstruct molecular phylogenetic relationships? The answer is to proceed hierarchically. The first step comprises sequence selection and alignment, to determine site-by-site homologies and to detect DNA or amino acid differences. The second step is to build a mathematical model describing the evolution in time of the sequences. A model can be built empirically, using properties calculated through comparisons of observed sequences, or parametrically, using chemical or biological properties of DNA and amino acids, as for instance hydrophobicity values of each amino acid. Such models permit estimation of the genetic distance between two homologous sequences, measured by the expected number of nucleotide substitutions per site that have occurred on the evolutionary lineages between them and their most recent common ancestor. Such distances may be represented as branch lengths in a phylogenetic tree; the extant sequences form the tips of the tree, whereas the ancestral sequences form the internal nodes and are generally not known.

The third step involves applying an appropriate statistical method to find the tree topology and branch lengths that best describe the sequences' phylogenetic relationships. The fourth step consists of the interpretation of results. We focus primarily on the second step. The statistical comparison of mathematical models of sequence evolution is in itself interesting, as the rejection of a simpler model in favor of one that incorporates additional hypothesized biological factors implies the real significance of those factors. Equally importantly, there is a growing body of evidence that phylogenetic inferences are more reliable the more accurate the model of sequence evolution, and this is another motivation for finding and using the best available models.

We proceed in this review by introducing some of the DNA base substitution and amino acid replacement models most widely used, along with some of their most important current developments. Maximum likelihood inference using these models is briefly introduced as one statistical method used to draw evolutionary inferences and as a fundamental part of the procedure for comparing the models. We conclude with speculations on the future course of molecular evolutionary analyses.

Models of Molecular Evolution

Introduction

In 1965, Zuckerkandl and Pauling (1965) proposed the theory of a molecular clock, that is, that the

[1]Corresponding author.
E-MAIL N.Goldman@gen.cam.ac.uk; FAX 01223-333992.

rate of molecular evolution is approximately constant over time for all the proteins in all lineages. According to this theory, any time of divergence between genes, proteins, or lineages can be dated simply by measuring the number of changes between sequences. Soon afterwards, in 1969, Jukes and Cantor (1969) proposed a stochastic model for DNA substitution in which all nucleotide substitutions occur at an equal rate, and when a nucleotide is substituted, any one of the other nucleotides is equally likely to be its replacement.

It is not surprising, however, that the molecular clock hypothesis and the Jukes and Cantor substitution model have both been found to be overly simple. Mutation rates seem to vary both among and within genomes, being affected by many factors such as chromosomal position (Sharp et al. 1989), G + C content (Wolfe 1991), nearest neighbor bases (Blake et al. 1992), and different efficiency of the repair systems between the lagging and the leading DNA strands during replication and transcription (Veaute and Fuchs 1993). Thus, any molecular clock seems to tick at different rates for different DNA positions. Furthermore, it is known that misincorporation errors during DNA duplication or repair are facilitated if a base is replaced by a similar one, and thus, transitions occur more frequently than transversions (Brown and Simpson 1982): often twice as frequently, but the ratio can be much higher. Differences in mutation rate tend to decrease TA and CG dimers and to produce an excess of CT and TG dimers (Ohno 1988). Sueoka introduced a theoretical model to explain the large variations in G + C content shown by the DNA sequences belonging to different species; he proposed a directional mutational pressure attributable to misincorporation errors during DNA repair or replication as a cause of such a variation (Sueoka 1992). The genomes of higher vertebrates and plants are mosaics of large regions (isochores) with remarkably different G + C content. Moreover, warm-blooded vertebrates seem to have higher mutation rates than cold-blooded vertebrates.

It is difficult to implement, in a model that aims to be general, all the different mutation rules and patterns that we detect in the genetic material belonging to different species. Instead, the models that have been used have incorporated only the simplest rules or have been based on empirical observation with little understanding of the underlying biology. We start our discussion of these models with a description of an assumption they all share, the Markov property.

Markov Models

Consider a stochastic model for DNA or amino acid sequence evolution. We assume independence of evolution at different sequence sites and thus can consider sites one by one. At any single site, the model works with probabilities $P_{ij}(T)$ that base i will have changed to base j after a time T. The subscripts i and j take the values 1,...,4 to represent the nucleotides A, T, C, G for DNA sequences and 1,...,20 for amino acid sequences.

Given a stochastic variable $X(t)$ describing the evolution through time t of a site in one sequence, the Markov assumption asserts that $P_{ij}(T) = \Pr[X(s + T) = j|X(s) = i]$ is independent of $s \geqslant 0$. Informally, this means that subsequent to any time s it does not matter how the process reached state i by time s (the process is ''memoryless''), and the future course of evolution depends only on i.

The probabilities of transition from one base to another, $P_{ij}(T)$, can be written as a matrix $\mathbf{P}(T)$, and then we can write

$$\mathbf{P}(T + dT) = \mathbf{P}(T)(\mathbf{I} + \mathbf{Q}dT)$$

where dT represents a small time, and $\mathbf{I}$ is the identity matrix. The matrix $\mathbf{Q}$ is known as the instantaneous rate matrix and has off-diagonal entries Q_{ij} equal to the rates of replacement of i by j. (The diagonal entries, Q_{ii}, are defined by a mathematical requirement that the row sums are all zero.) This equation is solved to give

$$\mathbf{P}(T) = e^{T\mathbf{Q}} = \mathbf{I} + T\mathbf{Q} + \frac{(T\mathbf{Q})^2}{2!} + \frac{(T\mathbf{Q})^3}{3!} + \ldots$$

Spectral decomposition (also termed diagonalization) of $\mathbf{Q}$ allows us to calculate the matrix $\mathbf{P}(T)$:

$$\mathbf{P}(T) = \mathbf{U} \cdot \text{diag}\{e^{\lambda_1 T}, \ldots, e^{\lambda_n T}\} \cdot \mathbf{U}^{-1}$$

where the matrix $\mathbf{U}$ contains the eigenvectors of $\mathbf{Q}$, the λ_i are the eigenvalues of $\mathbf{Q}$ and diag{} denotes the diagonal matrix of the elements contained in the braces. The components $P_{ij}(T)$ can be written as

$$P_{ij}(T) = \sum_k c_{ijk} e^{\lambda_k T}$$

where the sum is over k = 1,...,4 for DNA sequences and over k = 1,...,20 for amino acids; c_{ijk} is a function of $\mathbf{U}$ and $\mathbf{U}^{-1}$. Note that T and $\mathbf{Q}$ are confounded; $T\mathbf{Q} = (T/\gamma)(\gamma\mathbf{Q})$ for any $\gamma \neq 0$ (e.g., half the time at twice the rate has the same result). Therefore, absolute times T typically cannot be used, and in practice, time is scaled to units of expected substitutions per site.

A Markov process can have three important properties: homogeneity, stationarity, and reversibility. Homogeneity means that the rate matrix is independent of time, that is, that the patterns of nucleotide substitution or amino acid replacement remain the same in different parts of the tree. A homogeneous process has an equilibrium distribution that is also the limiting distribution when time approaches infinity. Stationarity means that the process is at that equilibrium, that is, nucleotide frequencies have remained more or less the same during the course of evolution. Reversibility means that $\pi_i P_{ij}(T) = \pi_j P_{ji}(T)$ for all i, j, and T where π_i are the frequencies of occurrence for each base. A consequence of reversibility is that the process of sequence evolution is theoretically indistinguishable from the same process watched in reverse.

Models in widespread use typically assume homogeneity, yet this is rarely likely to be fully appropriate, for example, because of the dependence of mutation on local sequence context. Stationarity is not a consequence of a Markov model but of its application; this too is generally assumed in phylogenetics, although when base frequencies are quite different in different species this assumption is clearly violated. Genomes show large differences in base compositions. For instance, the genome of the bacterium *Micrococcus luteus* has 74% G + C content, whereas the genome of the bacterium *Mycoplasma capricolum* has only 25% G + C content. Reversibility too is generally assumed, with little justification other than that numerical calculations are simplified considerably. Assumptions such as those of homogeneity, stationarity, and reversibility are typical of the approximations that have to be made to render our knowledge of molecular biology into a mathematically tractable form.

DNA Substitution Models

The model of Jukes and Cantor (1969) described above is defined by $Q_{ij} = \alpha$ for all i, $j = 1,\ldots,4$; $i \neq j$, meaning that each base is substituted by any other at equal rate α. A consequence of this model is that the base frequencies (π_i) are all assumed equal to 0.25. Kimura (1980) proposed a two-parameter model that considered the difference in transition and transversion rates. The instantaneous rate matrix can be written as

$$\mathbf{Q} = \begin{bmatrix} \cdot & \beta & \beta & \alpha \\ \beta & \cdot & \alpha & \beta \\ \beta & \alpha & \cdot & \beta \\ \alpha & \beta & \beta & \cdot \end{bmatrix}$$

For the sake of clarity, because the row sums of the matrix are constrained to equal zero, we have used dots on the diagonal. In this and all subsequent matrices, the order of the bases for columns and rows are A, T, C, G, and the (i,j) entry represents Q_{ij}, the rate ($i \rightarrow j$) at which a base i is replaced by a base j. After Kimura, several authors proposed models with increasing numbers of parameters. Blaisdell (1985) introduced an asymmetry for some reciprocal changes: $i \rightarrow j$ has a different substitution rate from $j \rightarrow i$:

$$\mathbf{Q} = \begin{bmatrix} \cdot & \gamma & \gamma & \alpha \\ \delta & \cdot & \alpha & \delta \\ \delta & \beta & \cdot & \delta \\ \beta & \gamma & \gamma & \cdot \end{bmatrix}$$

Unlike Kimura's two-parameter model, the four-parameter model proposed by Blaisdell does not have the property of reversibility. Further related contributions based respectively on four- and six-parameter models were made by Takahata and Kimura (1981) and Gojobori et al. (1982).

Felsenstein (1981) proposed a model in which the rate of substitution to a nucleotide depends only on the equilibrium frequency of that nucleotide. This approach adds three free parameters to the Jukes and Cantor (1969) model:

$$\mathbf{Q} = \begin{bmatrix} \cdot & \mu\pi_T & \mu\pi_C & \mu\pi_G \\ \mu\pi_A & \cdot & \mu\pi_C & \mu\pi_G \\ \mu\pi_A & \mu\pi_T & \cdot & \mu\pi_G \\ \mu\pi_A & \mu\pi_T & \mu\pi_C & \cdot \end{bmatrix}$$

Usually the nucleotide equilibrium frequencies can be estimated by simply analyzing base composition in the DNA sequences under study.

Hasegawa and coworkers (1985) implemented transition/transversion bias in Felsenstein's model, effectively combining it with the model of Kimura (1980). The rate matrix is

$$\mathbf{Q} = \begin{bmatrix} \cdot & \beta\pi_T & \beta\pi_C & \alpha\pi_G \\ \beta\pi_A & \cdot & \alpha\pi_C & \beta\pi_G \\ \beta\pi_A & \alpha\pi_T & \cdot & \beta\pi_G \\ \alpha\pi_A & \beta\pi_T & \beta\pi_C & \cdot \end{bmatrix}$$

Thus, with respect to Hasegawa's model, Kimura's model corresponds to the case $\pi_A = \pi_T = \pi_C = \pi_G = 0.25$; Felsenstein's model corresponds to the case of $\beta = \alpha$; and when both these simplifications are made, we obtain the model of Jukes and Cantor. Felsenstein (1995) and Tamura and Nei (1993) have also devised models very similar to that of Hasegawa et al. (1985).

The most general model can have at most 12

independent parameters; insisting on reversibility reduces this to 9. Such a model was considered by Tavaré (1986) and later by Yang (1994a) and can be parameterized as follows:

$$\boldsymbol{Q} = \begin{bmatrix} \cdot & \alpha\pi_T & \beta\pi_C & \gamma\pi_G \\ \alpha\pi_A & \cdot & \rho\pi_C & \sigma\pi_G \\ \beta\pi_A & \rho\pi_T & \cdot & \tau\pi_G \\ \gamma\pi_A & \sigma\pi_T & \tau\pi_C & \cdot \end{bmatrix}$$

The models described above are parametric, in the sense that they are defined in terms of parameters (π_i, α, β, etc.) inspired by our understanding of biology. Empirical models of nucleotide substitution have also been studied. These models are derived from the analysis of inferred substitutions in reference sequences, perhaps the sequences under current study or from databases. Advantages of this approach can be the better description of the evolution of the sequences under study, if a suitable reference set is used, particularly if this reference set is large. Disadvantages can be inaccuracy owing to an inappropriate reference set and a lack of a broader biological interpretability of purely empirical findings. These models have received less attention and use than parametric models. Examples of this approach are found in Lanave et al. (1984), Zharkikh (1994) and Arvestad and Bruno (1997).

All the models described so far operate at the level of individual nucleotides. In an attempt to introduce greater biological reality, through knowledge of the genetic code and the consequent effect of nucleotide substitutions in protein-coding sequences on the encoded amino acid sequences, Goldman and Yang (1994a) described a codon mutation model. They considered the 61 sense codons i consisting of nucleotides $i_1i_2i_3$. The rate matrix $\mathbf{Q}$ consisted of elements Q_{ij} describing the rate of change of codon $i = i_1i_2i_3$ to $j = j_1j_2j_3$ ($i \neq j$) depending on the number and type of differences between i_1 and j_1, i_2 and j_2, and i_3 and j_3 as follows:

$$Q_{ij} = \begin{cases} 0 & \text{if 2 or 3 of the pairs } i_k, j_k \text{ are different} \\ \mu\pi_j e^{-d_{aa_i,aa_j}/V} & \text{if one pair differs by a transversion} \\ \mu\kappa\pi_j e^{-d_{aa_i,aa_j}/V} & \text{if one pair differs by a transition} \end{cases}$$

where d_{aa_i,aa_j} is the distance between the amino acid coded by the codon i (aa_i) and the amino acid coded by the codon j (aa_j) as calculated by Grantham (1974) on the basis of the physicochemical properties of the amino acids. This model takes account of codon frequencies (through the π_j), transition/transversion bias (through κ), differences in amino acid properties between different codons (d_{aa_i,aa_j}), and levels of sequence variability (V). A similar model was described by Muse and Gaut (1994). Recent work by Yang and Nielsen (1998) and Pedersen et al. (1998) has developed and improved the Goldman and Yang (1994a) model.

Amino Acid Replacement Models

Base substitutions are more easily fixed in noncoding regions than in coding regions. In coding regions, natural selection determines the fixation of, for example, amino acid replacements or trinucleotide slippage, insertions, and deletions. Gene duplication events can also give rise to amino acid replacements, whereby one copy of a duplicated gene can accumulate a large number of mutations and acquire new substrate specialization. In addition, for distantly related sequences, the ''filtering'' of DNA sequences by the genetic code can give amino acid sequences with more obviously interpreted similarities, and amino acid sequences are less prone to have wide-ranging differences in composition (e.g., G + C richness) than are some DNA sequences. For these reasons, it can be valuable to look at models of amino acid replacement.

In contrast to DNA substitution models, amino acid replacement models have concentrated on the empirical approach. Dayhoff and coworkers (1972, 1978) developed a model of protein evolution that resulted in the development of a set of widely used replacement matrices. In the Dayhoff approach, replacement rates are derived from alignments of protein sequences that are at least 85% identical; this constraint ensures that the likelihood of a particular mutation (e.g., L $\rightarrow$ V) being the result of a set of successive mutations (e.g., L $\rightarrow x \rightarrow y \rightarrow$ V) is low. An implicit instantaneous rate matrix was estimated, and replacement probability matrices $\mathbf{P}(T)$ were generated for different values of T. One of the main uses of the Dayhoff matrices has been in database search methods where, for example, the matrices $\mathbf{P}(0.5)$, $\mathbf{P}(1)$, and $\mathbf{P}(2.5)$ (known as the PAM50, PAM100, and PAM250 matrices) are used to assess the significance of proposed matches between target and database sequences. However, the implicit rate matrix has been used for phylogenetic applications.

Recently, Jones et al. (1992) and Gonnet et al. (1992) have used much the same methodology as Dayhoff but with modern databases. The Jones et al. model has been implemented for phylogenetic analyses with some success. Jones et al. (1994) have also calculated an amino acid replacement matrix

specifically for membrane spanning segments. This matrix has remarkably different values from the Dayhoff matrices, which are known to be biased toward water-soluble globular proteins.

Adachi and Hasegawa (1995, 1996) have implemented a general reversible Markov model of amino acid replacement that uses a matrix derived from the inferred replacements in mitochondrial proteins of 20 vertebrate species. The investigators show that this model performs better than others when dealing with mitochondrial protein phylogeny.

A simple, nonempirical model of amino acid replacement was proposed by Nei (1987). This model implements a Poisson distribution and gives accurate estimates of the number of amino acid replacements when species are closely related.

A different approach was used by Henikoff and Henikoff (1992). They used local, ungapped alignments of distantly related sequences to derive the BLOSUM series of matrices. Matrices of this series are identified by a number after the matrix (e.g., BLOSUM50), which refers to the minimum percentage identity of the blocks of multiple aligned amino acids used to construct the matrix. It is noteworthy that these matrices are directly calculated without extrapolations, and are analogous to transition probability matrices $\mathbf{P}(T)$ for different values of T, estimated without reference to any rate matrix $\mathbf{Q}$. The BLOSUM matrices often perform better than PAM matrices for local similarity searches but have not been widely used in phylogenetics.

Rate Heterogeneity

One of the most important recent advances in the reconstruction of evolutionary trees is the consideration of heterogeneity of evolutionary rates among sites. The biological basis of heterogeneous mutation rate among sites probably reflects the influence of the nearest neighbors on mutation rate. Stacking energies along the molecule, helix configuration (A, B, Z-DNA, triple helix), supercoiling, and DNA intrinsic curvature (that is sequence dependent) change the solvent accessibility and thus base reactivity. The fixation of any mutation depends on DNA and protein structure/function selection pressures. Protein coding and noncoding DNA regions show remarkably different mutation rates; moreover, each codon position is subject to different selection pressures. The incorporation of heterogeneity of evolutionary rates among sites has led to a new set of models that generally provides a better fit to observed data, and phylogeny reconstruction has improved (e.g., Yang 1994b, 1996a; Yang et al. 1994).

Some authors have considered models in which a fraction of sites change at one rate, whereas the other sites are invariable (e.g., Hasegawa et al. 1985). More popular and successful have been models based on a continuous distribution of rates. Nei and Gojobori (1986), Yang (1993) and Tamura and Nei (1993) have modelled site rates using a Gamma distribution. A continuous distribution in which every site may have a different rate seems to be the most biologically plausible model. Yang, however, went on to show that the ''discrete Gamma model,'' with as few as four categories of evolutionary rates chosen to approximate a Gamma distribution, performs very well (Yang 1994b). It is also considerably more practical computationally.

The success of the Gamma distribution seems to be in its flexibility. With this model, we assume that the rate of substitution for each site is drawn from a Gamma distribution with shape parameter α. If α is <1, the distribution implies that there is a relatively large amount of rate variation, with many sites evolving very slowly but some sites evolving at a high rate. For values of $\alpha>1$, the shape of the distribution changes qualitatively, with less variation and most sites having roughly similar rates. It appears that the range of distributional shapes available under the permitted values of $0 < \alpha < \infty$ is well able to describe the variation found in DNA sequences.

Yang (1995) and Felsenstein and Churchill (1996) have implemented methods in which several categories of evolutionary rates can be defined. Both methods use hidden Markov model (HMM) techniques (see Rabiner 1989; Eddy 1996 and references therein) to describe the organization of areas of unequal and unknown rates at different sites along sequences. All possible assignments of evolutionary rate category at each site contributes to the phylogenetic analysis of sequences, and algorithms are also available to infer the most probable rate category for each site. These methods are not yet widely used. They are, however, among the first to consider the organization of sites along sequences instead of assuming that all sites evolve according to identical processes and independently of one another and so deserve more consideration in the future.

Combined Models

Cao et al. (1994) and Yang (1996b) have considered the problem of building models to analyze the sequences of multiple genes from the same set of species. Different sequences, such as protein coding regions, noncoding regions and tRNA genes, or sim-

ply different regions of the same gene, may show heterogeneity in their evolutionary processes. Parameters that might exhibit such heterogeneity are nucleotide frequencies, transition/transversion rate biases, and the extent of rate variation across sites. Combining heterogeneous data this way, using models appropriate to each part of the combined data, can give more powerful analyses. For example, Yang (1996b) considered tests of the molecular clock hypothesis. He illustrated a case in which the hypothesis of rate constancy among lineages could not be rejected when considering a number of genes singularly, but for which the combined set of sequences could indicate significant differences in substitution rates among species.

The Impact of Structural Biology

Recent phylogenetic analyses of DNA and protein sequences have been improved by incorporating structural and functional properties into inferential models. At the same time, phylogenetic relationships can also suggest additional clues to RNA and protein structure. A first approach is to consider information only indirectly related to structure, such as DNA G + C content (Churchill 1989) or physicochemical properties of amino acids, for instance, hydrophobicity, charge, and size (Naylor and Brown 1997). Churchill incorporated the local composition heterogeneity of DNA sequences as states of a hidden Markov chain. Local G + C content influences the structure and curvature of a DNA molecule, and Churchill's model can be regarded as the first to incorporate structural information.

Rzhetsky (1995) introduced a model to estimate base substitution in ribosomal RNA genes and to infer phylogenetic relationships. Phylogenetic analyses of ribosomal RNA (rRNA) sequences have given important results about ancient events because of their high levels of conservation over extremely long evolutionary times. Some authors have, however, suggested that ribosomal RNA trees may sometimes be misleading, especially when G + C content differs widely among lineages. Rzhetsky's model takes into account rRNA secondary structure elements, namely, stem and loop regions. Although it is very difficult to infer rRNA tertiary structure, fairly reliable predictions of stable folded secondary structures can be made. Stem and loop regions have different rates of base substitution: stems are double-stranded regions, and loops are single-strand RNA, and thus, selection pressures on them are different. Rzhetsky (1995) modelled stem regions using 16 discrete states (representing all the possible pairings) and loops using a 4-state rate matrix. For stems, the 240 off-diagonal elements of the rate substitution matrix each take one of four different values, depending on the type of substituting pairs; for loops, the simple Jukes and Cantor model is used.

Other important contributions are those of Vawter and Brown (1993), Schöniger and von Haeseler (1994), and Tillier and Collins (1998). Additional constraints and difficulties yet to be modelled for RNA secondary structure include G · U pairings and loop size; additionally, in tRNA, 10% of the nucleotides are rare variants. The step from secondary to tertiary structure will be yet more difficult to implement in evolutionary models.

Goldman, Thorne, and coworkers have introduced an evolutionary model that combines protein secondary structure and amino acid replacement (Goldman et al. 1996, 1998; Thorne et al. 1996; Liò et al. 1998). Their approach is related to that of Dayhoff and coworkers but considers different categories of structural environment, for example, α-helix, β-sheet, turn, and loop, with each category further classified by whether it is exposed to solvent or is buried in the protein core. Whereas the Dayhoff approach simply considers the ''average'' environment for each amino acid, Goldman, Thorne, and coworkers inferred a Markov chain model of amino acid replacement for each of the different categories. Underlying (but unobserved) transitions between the different categories along a protein-coding sequence are described with a (hidden) Markov chain. The resulting HMM allows the simultaneous inference of phylogeny and protein structure, using information about each to improve inference of the other.

Alignment, Phylogeny, and Evolution

The presence of multiple deletions, insertions, and repeats can make the alignment procedure a delicate passage in phylogeny building. Most phylogenetic studies are based on a fixed alignment, derived in advance of phylogenetic analysis and subsequently assumed to be correct. It is beyond the scope of this review to discuss these alignment techniques. With respect to evolution, however, we refer readers to the work of Thorne et al. (1991, 1992) and Mitchison and Durbin (1995) who have devised evolutionary models that simultaneously account for nucleotide or amino acid replacements and for processes of insertion and deletion.

Is it Time for Whole Genome–Based Phylogeny?

At the time when few genes had been sequenced, incongruences between biological sequence-based

phylogeny and morphology-based phylogeny fed the strong faith that the sequencing of larger amounts of genetic material could result in the complete banishment of doubts and incongruences. This has turned out not to be completely true. Cao et al. (1994) have shown that phylogeny based on different mitochondrial proteins can suggest wrong trees. Nowadays, the availability of entire genome or chromosome sequences points to the open questions of the choice of ''representative'' sets of genes and how phylogenies, based on different sets of genes with different evolutionary rate, can be combined to draw a single evolutionary landscape. Yang (1996b) has suggested the use of combined models for different genes; Cao et al. (1994) have shown that chaining the amino acid sequences of different proteins can lead to better (but not always completely satisfactory) solutions.

Pitfalls in phylogeny building can also depend on the presence of large families of orthologous and paralogous genes. The complete sequencing of genomes can lead to a better knowledge of paralogy events and thus to more robust phylogeny analyses. We feel it is still necessary to improve our ''small scale'' understanding through models such as those reviewed here in order to describe adequately genome-sized effects.

Applications

Introduction

The main aim of this review is to describe models currently used to describe sequence evolution. It is beyond our scope to include any comprehensive discussion of all their possible applications for phylogenetic estimation. Instead, we will concentrate on some of their applications through maximum likelihood (ML) methods, as we feel these methods provide the best framework that allows both phylogenetic estimation and the testing and comparison of alternative models that permit greater understanding of the processes of evolution in addition to its reconstruction.

Part of the appeal of ML methods lies in their robust mathematical and statistical basis and their ability to use all the available data. Recent studies based on simulations have led to the conclusions that ML methods are often more accurate in inferring the correct tree than other methods (Kuhner and Felsenstein 1994; Huelsenbeck 1995). The disadvantage of this method is the need for powerful computing resources. The number of possible unrooted bifurcating trees for n sequences is $(2n - 3)!! = 1 \times 3 \times \ldots \times (2n - 3)$ (Edwards and Cavalli-Sforza 1964; Felsenstein 1978); as n increases, this grows as fast as $n!! \times 2^n$. With $n = 10$ sequences the number of candidate trees exceeds 2×10^6, and the task of determining which is the ML tree can be likened to finding a one-half-gram needle in a 10^3-kg haystack. Various heuristic methods have been proposed to help overcome this difficulty (e.g., see Strimmer and von Haeseler 1996; Swofford et al. 1996), and there are a number of practical computer programs available. Some of those that we find most useful, or most interesting, are listed in Table 1.

Maximum Likelihood Estimation

The likelihood of a hypothesis is defined as the probability of the data given that hypothesis. In phylogeny reconstruction, the evolutionary tree (its shape and branch lengths) and any other free parameters of the evolutionary model represent these hypotheses. Different hypotheses have different likelihood values. ML evaluates competing hypotheses (trees and parameters) by selecting those with the highest likelihood, as it is these that render the observed data most plausible. Likelihood calcula-

Table 1. A Selection of Software Packages Useful for the Molecular Sequence Analyses Described in the Text

Software	WWW address	Reference
MOLPHY	anonymous FTP at sunmh.ism.ac.jp/pub/molphy	Adachi and Hasegawa (1995)
PAML	http://abacus.gene.ucl.ac.uk/ziheng/paml.html	Yang (1997a,b)
PASSML	http://ng-dec1.gen.cam.ac.uk/hmm/Passml.html	Liò et al. (1998)
PAUP*	http://chee.unm.edu/paup	Swofford (1998)
PHYLIP	http ://evolution.genetics.washington.edu/phylip.html	Felsenstein (1995)
PUZZLE	http://www.zi.biologie.uni-muenchen.de/~strimmer/puzzle.html	Strimmer and von Haeseler (1996)

tions for evolutionary trees are not straightforward, typically requiring computations that allow for all the possible unobserved sequences at the internal (ancestral) nodes of hypothesized trees. These calculations have, however, been possible in practice since the work of Felsenstein (1981). As well as estimating trees and parameters by those giving the highest likelihood scores, the likelihood values themselves are used in model comparisons.

Comparison of Models

Statistical tests for phylogenetic trees allow the estimation of the reliability of inferred trees. For example, the bootstrap technique introduced to phylogenetics by Felsenstein (1985) measures how well a group is reflected by all the data in a sequence alignment, given the data analysis method used.

More recently, tests and comparisons of the models used to describe sequence evolution have become available (Goldman 1993). Yang et al. (1994) show how two models that differ in only one parameter can be compared, by examining the estimated value and variance of that parameter. More generally, comparisons of different evolutionary models require consideration of their different numbers of parameters. One solution that has been proposed is the AIC (Akaike information criterion) test: a model that minimizes AIC = [$-2 \cdot$ log(likelihood)] + [2 $\cdot$(number of free parameters)] is considered to be the most appropriate (Akaike 1974; Kishino et al. 1990). The AIC has not been widely used in recent years, as likelihood ratio tests (LRTs) have increased in usage.

LRTs are a class of powerful statistical tests that compare the ML values of competing hypotheses. Statistical theory states that under ideal conditions these tests will have easily found properties based on χ^2 statistics. LRTs have found various applications in phylogenetics. Yang et al. (1995) implemented an LRT [first proposed by Felsenstein (1981)] of the molecular clock hypothesis and successfully used the predicted χ^2 distribution. Some other tests of closely related models are also found to be equally straightforward. However, the χ^2 distribution is not always useful, as found by Goldman (1993) when implementing an LRT proposed by Navidi et al. (1991) and Goldman (1993) of the overall goodness of fit of models of sequence evolution. Goldman (1993) has described a Monte Carlo simulation approach that can be used instead in such cases.

Results of Model Comparisons

Yang (1996c, 1997a,b) has recently reported a large comparison of models using a set of 895-bp mtDNA sequences from human, chimpanzee, gorilla, orangutan, and gibbon. These sequences have been analyzed previously by many other authors and can be regarded as a useful benchmark of the goodness of fit of nucleotide substitution models (Brown and Simpson 1982).

Simple models such as those of Jukes and Cantor (1969) and Kimura (1980), even when recovering the correct tree topology, can result in severe underestimation of branch lengths. At the same time, simple models can discriminate better between candidate trees (Goldman and Yang 1994b). In this respect and also with respect to components of other, more complex models (below), it remains unclear how best to accomodate the trade-off between incorporating into models enough complexity (or biological reality) to capture evolutionary information accurately and avoiding overparameterization that can lead to a loss of discriminatory power. It is our belief, however, that under biologically realistic conditions models that are amongst the most complex currently available will probably be most successful.

Yang (1994a) showed that for large data sets the general reversible model performs better than other models and the use of nonreversible models is not worthwhile. Goldman and Yang (1994a) showed that their codon-based model of protein-coding DNA sequences could fit data better than models based on individual nucleotides.

Yang et al. (1994) have compared several models with and without the assumption of a Gamma distribution of rates over sites using different sets of sequences. They found that the incorporation of heterogeneity of rates over sites into the model of Hasegawa et al. (1985) performed better than the other models tested, in terms of fit of model to data and also the accuracy of estimated trees. Heterogeneity in rate variation over nucleotide sites can be the most important parameter determining the goodness of fit of a model (Yang et al. 1994); in particular, the incorporation of such a parameter helps avoid the severe underestimation of long branch lengths that can occur with other methods. At the same time, misclassification of rates at sites can affect likelihoods and thus tree topology estimation (Yang et al. 1994).

Goldman, Thorne, and coworkers used simulation techniques to evaluate the worth of incorporating structural information into amino acid replacement models, relative to other models that do not consider such information (e.g., Goldman et al. 1998). They showed that the incorporation of sol-

vent accessibility at each site results in a conspicuous improvement of the likelihood and that the incorporation of secondary structure environment (e.g., α-helix, β-sheet, turn, loop) at each site results in a further significant improvement. Incorporating knowledge of typical length distributions of the secondary structure categories was not always significant (Goldman et al. 1998). This result leaves open the question of whether currently manageable data sets contain enough information to tell us more about evolutionary replacement processes or whether a secondary structure-based model with fewer parameters is equally able to catch all the available information.

Perspectives and Conclusion

Phylogenetic studies are probably going to be increasingly based on structural biological data and on statistical formalization. This reflects the importance of improved models and of extracting the maximum information from sequence data. In addition, the understanding of correlations or discrepancies between molecular and morphological evolution is a new and challenging frontier (e.g., Pagel 1994; Omland 1997), although one in which progress is perhaps slower.

Studies incorporating structural information are quite fruitful at the moment. As well as the work described above, additional complexity is gradually being introduced. For example, three-dimensional structure is being considered, in methods that look for correlations in sequence evolution (D.D. Pollock, W.R. Taylor, and N. Goldman, in prep.). Correlations between sites distant in the linear sequence of a protein will often reflect effects on parts of the protein that are very close in the folded (three-dimensional) structure. As such analyses become more specialized, however, there is some concern over whether there will ever be enough data to find these correlations reliably.

Regarding statistical formalization, current research is successfully generalizing the statistical testing of phylogenetic hypotheses. Huelsenbeck and Rannala (1997) have recently discussed likelihood ratio testing in phylogenetics from a practical and biologist-oriented point of view. They have stressed that both the use of the χ^2 distribution and Monte Carlo simulations have become indispensable tools for biologists dealing with competing phylogenetic hypotheses.

With our improving ability to handle phylogenetic questions statistically has come an increased interest in experimental design in phylogenetics. Much interest has concentrated on the estimation of large phylogenies [see the recent special issue of *Systematic Biology* **47**(1), 1998]. In addition, new methods have been described that allow more general questions (e.g., How valuable is it to add more species to a tree with a certain topology? Is it going to make the tree topology more robust if I sequence additional DNA regions? Which gene region should I sequence to get reliable phylogeny estimates?) to be answered (Goldman 1998; Graybeal 1998; Yang 1998).

Another recent methodological advance has been the introduction of modern computational statistical methods into phylogenetics. This has included renewed interest in Bayesian statistics and the use of sampling techniques such as Markov chain Monte Carlo (see Gilks et al. 1996 and papers therein), the realistic application of both of which have become possible only with the advent of modern computers. Bayesian approaches were used previously in linkage analysis but did not initially elicit particular attention, perhaps because of biologists' frequency-oriented background and attitude. In phylogenetics, however, there has been more interest. Yang and Rannala (1997) have recently used Bayesian methods for estimating phylogenetic trees. They used birth–death process models of ancestral speciation and extinction to specify the prior distribution of phylogenies and a Markov chain Monte Carlo method to estimate posterior probabilities of trees. Similar techniques were used by Mau et al. (1998) to generate confidence sets of phylogenetic trees.

In conclusion, we are delighted to report that the modelling of processes of sequence evolution is a thriving field of research. It has two immediate and important benefits: the improved understanding of the biological processes that shape evolution at the molecular level and the improved ability to infer from sequence data the story of the evolution of life on earth.

ACKNOWLEDGMENTS

P.L. is supported by an Engineering and Physical Sciences Research Council and Biotechnology and Biological Science Research Council Bioinformatics Joint Initiative grant; N.G. is supported by a Wellcome Trust Fellowship in Biodiversity Research.

REFERENCES

Adachi, J. and M. Hasegawa. 1995. *MOLPHY: Programs for molecular phylogenetics Ver. 2.3*. Institute of Statistical Mathematics, Tokyo, Japan.

———. 1996. Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J. Mol. Evol.* **42:** 459–468.

Akaike, H. 1974. Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information Theory* (ed. B.N. Petrov and F. Csaki), pp. 267–281. Akademia Kiado, Budapest, Hungary.

Arvestad, L. and W. Bruno. 1997. Estimation of reversible substitution matrices from multiple pairs of sequences. *J. Mol. Evol.* **45:** 696–703.

Blaisdell, B.E. 1985. A method for estimating from two aligned present day DNA sequences their ancestral composition and subsequent rates of composition and subsequent rates of substitution, possibly different in the two lineages, corrected for multiple and parallel substitutions at the same site. *J. Mol. Evol.* **22:** 69–81.

Blake, R.D., T.H. Samuel, and J. Nicholson-Tuell. 1992. The influence of nearest neighbors on rate and pattern of spontaneous point mutations. *J. Mol. Evol.* **34:** 189–200.

Brown, G.G. and M.V. Simpson. 1982. Novel features of animal mtDNA evolution as shown by sequences of two rat cytochrome oxidase subunit II genes. *Proc. Natl. Acad. Sci.* **79:** 3246–3250.

Cao, Y., J. Adachi, A. Janke, S. Pääbo, and M. Hasegawa. 1994. Phylogenetic relationships among Eutherian orders estimated from inferred sequences of mitochondrial proteins: Instability of a tree based on a single gene. *J. Mol. Evol.* **39:** 519–527.

Churchill, G.A. 1989. Stochastic models for heterogeneous DNA sequences. *Bull. Math. Biol.* **51:** 79–94.

Dayhoff, M.O., R.V. Eck, and C.M. Park. 1972. A model of evolutionary change in proteins. In *Atlas of protein sequence and structure vol. 5* (ed. M.O. Dayhoff), pp. 89–99. National Biomedical Research Foundation, Washington, DC.

Dayhoff, M.O., R.M. Schwartz, and B.C. Orcutt. 1978. A model of evolutionary change in proteins. In *Atlas of protein sequence and structure vol. 5 suppl. 2* (ed. M.O. Dayhoff), pp. 345–352. National Biomedical Research Foundation, Washington, DC.

Eddy, S. 1996. Hidden Markov models. *Curr. Opin. Struct. Biol.* **6:** 361–365.

Edwards, A.W.F. and L.L. Cavalli-Sforza. 1964. Reconstrucion of evolutionary trees. In *Phenetic and phylogenetic classification* (ed. V.H. Heywood and J. McNeill), pp. 67–76. Systematics Association, London, UK.

Felsenstein, J. 1978. The number of evolutionary trees. *Syst. Zool.* **22:** 240–249.

———. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **17:** 368–376.

———. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39:** 783–791.

———. 1995. PHYLIP (Phylogenetic inference package) ver. 3.57. Department of Genetics, University of Washington, Seattle, WA.

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

Gilks, W.R., S. Richardson, and D.J. Spiegelhalter, ed. 1996. *Markov chain Monte Carlo in practice*. Chapman and Hall, London, UK.

Gojobori, T., K. Ishii, and M. Nei. 1982. Estimation of the average number of nucleotide substitutions when the rate of substitution varies with nucleotide. *J. Mol. Evol.* **18:** 414–423.

Goldman, N. 1993. Statistical tests of models of DNA substitution. *J. Mol. Evol.* **36:** 182–198.

———. 1998. Phylogenetic information and experimental design in molecular systematics. *Proc. R. Soc. Lond. B* **265:** 1799–1786.

Goldman, N. and Z. Yang. 1994a. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11:** 725–736.

———. 1994b. Models of DNA substitution and the discrimination of evolutionary parameters. In *Proceedings of the XVIIth International Biometrics Conference, Hamilton, Ontario, Canada, vol. 1: Invited papers*, pp. 407–421. International Biometric Society, Hamilton, Ontario, Canada.

Goldman, N., J.L. Thorne, and D.T. Jones. 1996. Using evolutionary trees in protein secondary structure prediction and other comparative sequence analyses. *J. Mol. Biol.* **263:** 196–208.

———. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* **149:** 445–458.

Gonnet, G.H., M.A. Cohen, and S.A. Benner. 1992. Exaustive matching of the entire protein sequence database. *Science* **256:** 1443–1145.

Grantham, R. 1974. Amino acid difference formula to help explain protein evolution. *Science* **185:** 862–864.

Graybeal, A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? *Syst. Biol.* **47:** 9–17.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22:** 160–174.

Henikoff, S. and J.G. Henikoff. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci.* **89:** 10915–10919.

Huelsenbeck, J.P. 1995. The robustness of two phylogenetic methods: Four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining. *Mol. Biol. Evol.* **12:** 843–849.

Huelsenbeck, J.P. and B. Rannala. 1997. Phylogenetic methods come of age: Testing hypotheses in an evolutionary context. *Science* **276:** 227–232.

Jones, D.T., W.R. Taylor, and J.M. Thornton. 1992. The rapid generation of mutation data matrices from protein sequences. *Comp. Appl. Biosci.* **8:** 275–282.

———. 1994. A mutation data matrix for transmembrane proteins. *FEBS Lett.* **339:** 269–275.

Jukes, T.H. and C.R. Cantor. 1969. Evolution of protein molecules. In *Mammalian protein metabolism* (ed. H.N. Munro), pp. 21–132. Academic Press, New York, NY.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **6:** 111–120.

Kishino, H., T. Miyata, and M. Hasegawa. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J. Mol. Evol.* **31:** 151–160.

Kuhner, M.K. and J. Felsenstein. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* **11:** 459–468. (See also Erratum. *Mol. Biol. Evol.* **12:** 525 [1995].)

Lanave, C., G. Preparata, C. Saccone, and G. Serio. 1984. A new method for calculating evolutionary substitution rates. *J. Mol. Evol.* **20:** 86–93.

Liò, P., N. Goldman, J.L. Thorne, and D.T. Jones. 1998. PASSML: Combining evolutionary inference and protein secondary structure prediction. *Bioinformatics* **14:** 726–733.

Mau, B., M.A. Newton, and B. Larget. 1998. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics* (in press).

Mitchison, G. and R. Durbin. 1995. Tree-based maximum likelihood substitution matrices and hidden Markov models. *J. Mol. Evol.* **41:** 1139–1151.

Muse, S.V. and B.S. Gaut. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with applications to the chloroplast genome. *Mol. Biol. Evol.* **11:** 715–724.

Navidi, W.C., G.A. Churchill, and A. von Haeseler. 1991. Methods for inferring phylogenies from nucleic acid sequence data by using maximum likelihood and linear invariants. *Mol. Biol. Evol.* **8:** 128–143.

Naylor, G. and W.M. Brown. 1997. Structural biology and phylogenetic estimation. *Nature* **388:** 527–528.

Nei, M. 1987. *Molecular evolutionary genetics*. Columbia University Press, New York, NY.

Nei, M. and T. Gojobori. 1986. Simple methods for estimating the number of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3:** 418–426.

Ohno, S. 1988. Universal rule for coding sequence construction: TA/CG deficiency–TG/CT excess. *Proc. Natl. Acad. Sci.* **85:** 9630–9634.

Omland, K.E. 1997. Correlated rates of molecular and morphological evolution. *Evolution* **51:** 1381–1393.

Pagel, M. 1994. Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proc. R. Soc. Lond. B* **255:** 37–45.

Pedersen, A.-M.K., C. Wiuf, and F.B. Christiansen. 1998. A codon-based model designed to describe lentiviral evolution. *Mol. Biol. Evol.* **15:** 1069–1081.

Rabiner, L.R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77:** 257–286.

Rzhetsky, A. 1995. Estimating substitution rates in ribosomal RNA genes. *Genetics* **141:** 771–783.

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

Sharp, P.M., D.C. Shields, K.H. Wolfe, and W.-H. Li. 1989. Chromosomal location and evolutionary rate variation in enterobacterial genes. *Science* **258:** 808–810.

Strimmer, K. and A. von Haeseler. 1996. Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* **13:** 964–969.

Sueoka, N. 1992. Directional mutational pressure: Selective constraints and genetic equilibria. *J. Mol. Evol.* **34:** 95–99.

Swofford, D.L. 1998. *PAUP* 4.0: *Phylogenetic analysis using parsimony (and other methods)*. Sinauer Associates, Inc., Sunderland, MA.

Swofford, D.L., G. Olsen, P.J. Waddell, and D.M. Hillis. 1996. Phylogenetic inference. In *Molecular systematics*, 2nd ed. (ed. D.M. Hillis, C. Moritz, and B.K. Mable), pp. 407–514. Sinauer Associates, Inc., Sunderland, MA.

Takahata, N. and M. Kimura. 1981. A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes. *Genetics* **98:** 641–657.

Tamura, K. and M. Nei. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10:** 512–526.

Tavaré, S. 1986. Some probabilistic and statistical problems

in the analysis of DNA sequences. In *Lectures in mathematics in the life sciences*, Vol. 17 (ed. R.M. Miura), pp. 57–86. American Mathematical Society, Providence, RI.

Thorne, J.L., H. Kishino, and J. Felsenstein. 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.* **33:** 114–124. (See also Erratum. *J. Mol. Evol.* **34:** 91 [1992].)

———. 1992. Inching towards reality: An improved likelihood model of sequence evolution. *J. Mol. Evol.* **34:** 3–16.

Thorne, J.L., N. Goldman, and D.T. Jones. 1996. Combining protein evolution and secondary structure. *Mol. Biol. Evol.* **13:** 666–673.

Tillier, E.R. and R.A. Collins. 1998. High apparent rate of simultaneously compensatory base-pair substitutions in ribosomal RNA. *Genetics* **148:** 1993–2002.

Vawter, L. and W.M. Brown. 1993. Rates and patterns of base change in the small subunit ribosomal RNA genes. *Genetics* **134:** 597–608.

Veaute, X. and R. Fuchs. 1993. Greater susceptibility to mutations in lagging strands of DNA replication in *Escherichia coli* than in leading strand. *Science* **261:** 598–601.

Wolfe, K.H. 1991. Mammalian DNA replication: Mutation biases and the mutation rate. *J. Theor. Biol.* **149:** 441–451.

Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* **10:** 1396–1401.

———. 1994a. Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* **39:** 105–111.

———. 1994b. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J. Mol. Evol.* **39:** 306–314.

———. 1995. A space-time process model for the evolution of DNA sequences. *Genetics* **139:** 993–1005.

———. 1996a. Among-site rate variation and its impact on phylogenetic analysis. *Trends Ecol. Evol.* **11:** 367–372.

———. 1996b. Maximum-likelihood models for combined analyses of multiple sequence data. *J. Mol. Evol.* **42:** 587–596.

———. 1996c. Phylogenetic analysis using parsimony and likelihood methods. *J. Mol. Evol.* **42:** 294–307.

———. 1997a. *Phylogenetic analysis by maximum likelihood (PAML) ver. 1.3*. Department of Biology, University College London, UK.

———. 1997b. PAML: A program package for phylogenetic analysis by maximum likelihood. *Comm. Appl. Biosci.* **13:** 555–556.

———. 1998. On the best evolutionary rate for phylogenetic analysis. *Syst. Biol.* **47:** 125–133.

Yang, Z. and R. Nielsen. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J. Mol. Evol.* **46:** 409–418.

Yang, Z. and B. Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. *Mol. Biol. Evol.* **14:** 717–724.

Yang, Z., N. Goldman, and A. Friday. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol. Biol. Evol.* **11:** 316–324.

———. 1995. Maximum likelihood trees from DNA sequences: A peculiar statistical estimation problem. *Syst. Biol.* **44:** 384–399.

Zharkikh, A. 1994. Estimation of evolutionary distances between nucleotide sequences. *J. Mol. Evol.* **39:** 315–329.

Zuckerkandl, E. and L. Pauling. 1965. Evolutionary divergence and convergence in proteins. In *Evolving genes and proteins* (ed. V. Bryson and H. Vogel), pp. 97–166. Academic Press, New York, NY.