



mathematics of
evolution & phylogeny

EDITED BY
OLIVIER GASCUEL



MATHEMATICS OF EVOLUTION AND PHYLOGENY

This page intentionally left blank

Mathematics of Evolution and Phylogeny

Edited by

OLIVIER GASCUEL

OXFORD
UNIVERSITY PRESS

LIKELIHOOD CALCULATION IN MOLECULAR PHYLOGENETICS

David Bryant, Nicolas Galtier, and Marie-Anne Poursat

Likelihood estimation is central to many areas of the natural and physical sciences and has had a major impact on molecular phylogenetics. In this chapter we provide a concise review of some of the theoretical and computational aspects of likelihood-based phylogenetic inference. We outline the basic probabilistic model and likelihood computation algorithm, as well as extensions to more realistic models and strategies of likelihood optimization. We survey several of the theoretical underpinnings of the likelihood framework, reviewing research on consistency, identifiability, and the effect of model mis-specification, as well as advantages, and limitations, of likelihood ratio tests.

2.1 Introduction

Maximum likelihood (ML) estimation is arguably the most widely used method for statistical inference. The framework was introduced in the early 1920s by the pioneering statistician and geneticist, R.A. Fisher [18]. Likelihood based estimation is now routinely applied in almost all fields of the biological sciences, including epidemiology, ecology, population genetics, quantitative genetics, and evolutionary biology.

This chapter provides a concise survey of computational, statistical, and mathematical aspects of likelihood inference in phylogenetics. Readers looking for a general introduction to the area are encouraged to consult Felsenstein [15] or Swofford *et al.* [49]. A detailed mathematical treatment is provided by Semple and Steel [42].

Likelihood starts with a model of how the data arose. This model gives a probability $\mathbb{P}[D|\theta]$ of observing the data, given particular values for the parameters of the model (here denoted by the symbol θ). In phylogenetics, the parameters θ include the tree, branch lengths, the sequence evolution model, and so on. The key idea behind likelihood is to choose the parameters that maximize the probability of observing the data we have observed. We therefore define a *likelihood function* $L(\theta) = \mathbb{P}[D|\theta]$ (sometimes written as $L(\theta|D) = \mathbb{P}[D|\theta]$) that captures how “likely” it is to observe the data for a given value of the parameters θ . A high likelihood indicates a good fit. The *maximum likelihood estimate* is the value of

θ that maximizes $L(\theta)$. In our context, we will be searching for the maximum likelihood estimate of a phylogeny.

For the remainder of the chapter we will assume that the reader is comfortable with the concepts and terminology of likelihood in general statistics. Background material on likelihood (and related topics in statistics) can be found in Edwards [11] and Ewens and Grant [12].

Molecular phylogenetics is the field aiming at reconstructing evolutionary trees from DNA sequence data. The maximum likelihood (ML) method was introduced to this field by Joe Felsenstein [14] in 1981, and since become increasingly popular, particularly following recent increases in computing power.

Maximum likelihood has an important advantage over the still popular *maximum parsimony* (MP) method: ML is *statistically consistent* (see Section 2.6). As the size of the data set increases, ML will converge to the true tree with increasing certainty (provided, of course, that the model is sufficiently accurate). Felsenstein showed that Maximum Parsimony is *not* consistent, particularly in the case of unequal evolutionary rates between different lineages [13].

While the basic intuition behind likelihood inference is straightforward, the application of the framework is often quite difficult. First there is the problem of model design. In molecular phylogenetics, the evolution of genetic sequences is usually modelled as a Markov process running along the branches of a tree. The parameters of the model include the tree topology, branch lengths, and characteristics of the Markov process. As in all applied statistics there is a pay-off between more complex, realistic models, and simpler, tractable models. More complex models result in a better fit, but are more vulnerable to random error.

The second major difficulty with likelihood based inference is the problem of computing likelihood values and optimizing the parameters. Likelihood in molecular phylogenetics is made possible by the dynamic programming algorithm of Felsenstein [14]. We outline this algorithm in Section 2.3. However, nobody has found an efficient and exact algorithm for optimizing the parameters. The techniques most widely used are surprisingly basic.

The third difficulty with likelihood is the interpretation and validation of the results of a likelihood analysis: assessing which results are significant and which analyses are reliable.

In this chapter, we will discuss all three aspects. First (Section 2.2) we describe the basic Markov models central to likelihood inference in molecular phylogenetics. Second we present the fundamental algorithm of Felsenstein (Section 2.3), as well as extensions to more complex models (Section 2.4), and a survey of optimization techniques used (Section 2.5). Third we review the theoretical underpinnings of the likelihood framework. In particular, we discuss the consistency of maximum likelihood estimation in phylogenetics, and the conditions under which maximum likelihood will return the correct tree (Section 2.6). Finally, we show how the likelihood framework can guide us in the development of improved evolutionary models, and outline the theoretical justification for the standard likelihood ratio tests already in wide use in phylogenetics (Section 2.7).

2.2 Markov models of sequence evolution

Before any likelihood analysis can take place we need to formulate a probabilistic model for evolution. In reality, the process of evolution is so complex and multifaceted that there is no way we can completely determine accurate probabilities. Our descriptions of the basic model will involve assumption built upon assumption. It is a wonder of phylogenetics that we can get so far with the basic models that we do have. Of course, this phenomenon is in no way unique to phylogenetics.

The reliance of likelihood methods on explicit models is sometimes seen as a weakness of the likelihood framework. On the contrary, the need to make explicit assumptions is a strength of the approach. Likelihood methods enable both inferences about evolutionary history and assessments of the accuracy of the assumptions made. “The purpose of models is not to fit the data, but to sharpen the questions.”¹ While the basic models we describe in this section do an excellent job explaining much of the random variation in molecular sequences, shortcomings of the models (e.g. with respect to rate variation) have led to better models, a better understanding of sequence evolution, and a host of “sharper and sharper” questions on the relationship between rate variation, structure, and function.

More detailed reviews of these models can be found in references. [15, 49].

2.2.1 Independence of sites

Our first simplifying assumption is the perhaps unrealistic assertion that sites evolve independently. Thus the probability that sequence A evolves to sequence B equals the product, over all sites i , that the state in site i of A evolves to the state in site i of B . This simplifies computation substantially. In fact it is almost essential for tractability (though can be stretched a little—see Section 2.4). With this assumption made, we spend the rest of the section focusing on the evolution of an individual site.

2.2.2 Setting up the basic model

Consider the cartoon representation of site evolution in Fig. 2.1. Over a time period t , the state A at the site is replaced by the state T . There are a number of random mutation events (in this case, three) that are randomly distributed through the time period. One of these is redundant, with A being replaced by A . We consider these redundant mutations more for mathematical convenience than anything else. The mutations from A to G and from G to T are said to be *silent*. We do not observe the change to G , only the beginning and end states.

Let \mathcal{E} denote the set of states and let $c = |\mathcal{E}|$. For DNA sequences, $\mathcal{E} = \{A, C, G, T\}$, while for proteins, \mathcal{E} equals the set of amino acids. For convenience, we assume that the states have indices 1 to $|\mathcal{E}|$. The mutation events occur according to a *continuous time Markov chain* with state set \mathcal{E} . The number

¹Samuel Karlin, 11th R.A. Fisher Memorial Lecture, Royal Society 20, April 1983.

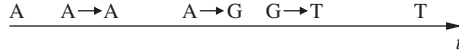


FIG. 2.1. Redundant and hidden mutations. Over time t , the site has a redundant mutation, followed by a mutation to G and then to T . The mutation to G is non-detectable, so is called *silent*. The timing (and number) of the mutation events is modelled by a Poisson process.

of these events has Poisson distribution: the probability of k mutation events is

$$\mathbb{P}[k \text{ events}] = \frac{(\mu t)^k e^{-\mu t}}{k!}.$$

Here μ is the rate of these events, so that the expected number of events in time t is μt . When there is a mutation event, we let \mathbf{R}_{xy} denote the probability of changing to state y given that the site was in state x . Since redundant mutations are allowed, $\mathbf{R}_{xx} > 0$. Putting everything together, the probability of ending in state y after time t given that the site started in state x is given by the xy th element of $\mathbf{P}(t)$, where $\mathbf{P}(t)$ is the matrix valued function

$$\mathbf{P}(t) = \sum_{k=0}^{\infty} (\mathbf{R}^k) \frac{(\mu t)^k e^{-\mu t}}{k!}. \quad (2.1)$$

This formula just expresses the probabilities of change summed over the possible values of k , the number of mutation events.

Let \mathbf{Q} be the matrix $\mathbf{R} - \mathbf{I}$, where \mathbf{I} is the $c \times c$ identity matrix. After some matrix algebra, equation (2.1) becomes

$$\mathbf{P}(t) = \sum_{k=0}^{\infty} \frac{(\mathbf{R} - \mathbf{I})^k (\mu t)^k}{k!} = \sum_{k=0}^{\infty} \frac{(\mathbf{Q} \mu t)^k}{k!} = e^{\mathbf{Q} \mu t}. \quad (2.2)$$

The matrix \mathbf{Q} is called the *instantaneous rate matrix* or *generator*. Here, $e^{\mathbf{Q} \mu t}$ denotes the *matrix exponential*. There is a standard trick to compute it.

First, diagonalize the matrix \mathbf{Q} as $\mathbf{Q} = \mathbf{A} \mathbf{D} \mathbf{A}^{-1}$ with \mathbf{D} diagonal (e.g. using Singular Value Decomposition, see [27]). For any integer k , we have that

$$\begin{aligned} (\mathbf{Q})^k &= (\mathbf{A} \mathbf{D} \mathbf{A}^{-1}) (\mathbf{A} \mathbf{D} \mathbf{A}^{-1}) \dots (\mathbf{A} \mathbf{D} \mathbf{A}^{-1}) \\ &= \mathbf{A} (\mathbf{D})^k \mathbf{A}^{-1}. \end{aligned}$$

Taking the powers of diagonal matrices is just a matter of taking the powers of its entries. It follows that

$$e^{\mathbf{Q} \mu t} = \mathbf{A} e^{\mathbf{D} \mu t} \mathbf{A}^{-1},$$

where $e^{\mathbf{D}}$ is a diagonal matrix and, for each x , $(e^{\mathbf{D}})_{xx} = e^{\mathbf{D}_{xx}}$.

As an example, consider the F81 model of Felsenstein [14]. We assume that the states in \mathcal{E} are ordered A, C, G, T . The model is defined in reference [49] in terms of its rate matrix

$$\mathbf{Q} = \begin{bmatrix} -(\pi_Y + \pi_G) & \pi_C & \pi_G & \pi_T \\ \pi_A & -(\pi_R + \pi_T) & \pi_G & \pi_T \\ \pi_A & \pi_C & -(\pi_Y + \pi_A) & \pi_T \\ \pi_A & \pi_C & \pi_G & -(\pi_R + \pi_C) \end{bmatrix}. \quad (2.3)$$

Rows in \mathbf{Q} indicate the initial state, and columns the final state, states being taken in the A, C, G, T alphabetic order. $\pi_A, \pi_C, \pi_G, \pi_T$ are probabilities that sum to one (see the next section), $\pi_R = \pi_A + \pi_G$ and $\pi_Y = \pi_C + \pi_T$. This model is equivalent to one with discrete generations occurring according to a Poisson process, and (single event) transition probability matrix

$$\mathbf{R} = \begin{bmatrix} 1 - (\pi_Y + \pi_G) & \pi_C & \pi_G & \pi_T \\ \pi_A & 1 - (\pi_R + \pi_T) & \pi_G & \pi_T \\ \pi_A & \pi_C & 1 - (\pi_Y + \pi_A) & \pi_T \\ \pi_A & \pi_C & \pi_G & 1 - (\pi_R + \pi_C) \end{bmatrix}.$$

The corresponding transition probability matrix, for a given time period t , is obtained by diagonalizing \mathbf{Q} and taking the exponential. The resulting matrix can be expressed simply by

$$\mathbf{P}_{xy}(t) = \begin{cases} \pi_y + (1 - \pi_y)e^{-\mu t}, & \text{if } x = y, \\ \pi_y(1 - e^{-\mu t}), & \text{if } x \neq y. \end{cases} \quad (2.4)$$

2.2.3 Stationary distribution

We have described here a *continuous time Markov chain*, the continuous time analogue of a *Markov chain*. We will also assume that this Markov process is *ergodic*. This means that as t goes to infinity, the probability that the site is in some state y is non-zero and independent of the starting state. That is, there are positive values π_1, \dots, π_c such that, for all x, y in EE

$$\lim_{t \rightarrow \infty} \mathbf{P}_{xy}(t) = \pi_y.$$

The values π_1, \dots, π_c comprise a *stationary distribution* (also called the *equilibrium distribution* or *equilibrium frequencies*) for the states. For all $t \geq 0$ these values satisfy

$$\pi_y = \sum_{x \in \mathcal{E}} \pi_x \mathbf{P}_{xy}(t). \quad (2.5)$$

If we sample the initial state from the stationary distribution, then run the process for time t , then the distribution of the final state will equal the stationary distribution. A consequence of equation (2.5) is that

$$0 = \sum_{x \in \mathcal{E}} \pi_x \mathbf{Q}_{xy},$$

so that we can recover the stationary distribution directly from \mathbf{Q} . We use $\mathbf{\Pi}$ to denote the $c \times c$ diagonal matrix with π_x 's down the diagonal.

For the F81 model we see from equation (2.4) that

$$\lim_{t \rightarrow \infty} \mathbf{P}_{xy}(t) = \pi_y,$$

for $x = y$ or $x \neq y$. The values $\pi_A, \pi_C, \pi_G, \pi_T$ make up the stationary distribution for this model. Hence π_R is the stationary probability for purines (A or G) and π_Y is the stationary probability for pyrimidines (C or T). The matrix $\mathbf{\Pi}$ is given by

$$\mathbf{\Pi} = \begin{bmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_C & 0 & 0 \\ 0 & 0 & \pi_G & 0 \\ 0 & 0 & 0 & \pi_T \end{bmatrix}.$$

2.2.4 Time reversibility

The next common assumption is of *time reversibility*. This is not exactly what it sounds like. We do not assume that the probability of going from state x to state y is the same as the probability of going from state y to state x . Instead we assume that the probability of sampling x from the stationary distribution and going to state y is the same as the probability of sampling y from the stationary distribution and going to state x . That is, for all $x, y \in \mathcal{E}$ and $t \geq 0$ we have

$$\pi_x \mathbf{P}_{xy}(t) = \pi_y \mathbf{P}_{yx}(t).$$

One can show that this corresponds to the condition that

$$\pi_x \mathbf{Q}_{xy} = \pi_y \mathbf{Q}_{yx},$$

that is, the matrix $\mathbf{\Pi Q}$ is symmetric.

The F81 model is time reversible even though $\mathbf{P}(t)$ is not symmetric. To see this, consider arbitrary states x, y with $x \neq y$. Then

$$\pi_x \mathbf{P}_{xy}(t) = \pi_x \pi_y (1 - e^{-\mu t}),$$

$$\pi_y \mathbf{P}_{yx}(t) = \pi_y \pi_x (1 - e^{-\mu t}).$$

Time reversibility makes it much easier to diagonalize \mathbf{Q} . Since $\mathbf{\Pi Q}$ is symmetric, so is

$$\mathbf{\Pi}^{-1/2} \mathbf{\Pi Q \Pi}^{-1/2} = \mathbf{\Pi}^{1/2} \mathbf{Q \Pi}^{-1/2}.$$

Finding eigenvalues of a symmetric matrix is, in general, far easier than finding eigenvalues of a non-symmetric matrix [27]. Hence we first diagonalize

$$\mathbf{\Pi}^{1/2} \mathbf{Q \Pi}^{-1/2}$$

to give a diagonal matrix \mathbf{D} and invertible matrix \mathbf{B} such that

$$\mathbf{\Pi}^{1/2} \mathbf{Q \Pi}^{-1/2} = \mathbf{B D B}^{-1}.$$

Setting $\mathbf{A} = \mathbf{\Pi}^{-1/2} \mathbf{B}$ gives $\mathbf{Q} = \mathbf{A D A}^{-1}$. This approach is used by David Swofford when computing the exponential matrices of general rate matrices in

PAUP [48]. Time reversibility also makes it easier to compute likelihoods on a tree, since the likelihood becomes independent of the position of the root [14].

2.2.5 Rate of mutation

In molecular phylogenetics, time is measured in *expected mutations per site*, rather than in years. The reason is that the rate of evolution can change markedly between different species, different genes, or even different parts of the same sequence.

Recall that our model of site evolution has mutation events occurring according to a Poisson process, with an expected number of events equal to μt . However, some of these mutation events are nothing more than mathematical conveniences—the mutations from a state to itself. If we assume that the distribution of the initial state equals the stationary distribution, then the probability that a mutation event gives a redundant mutation is

$$\sum_{x \in \mathcal{E}} \pi_x \mathbf{R}_{xx} = \text{trace}(\mathbf{\Pi R}).$$

Hence the probability that the mutation event is not redundant is

$$1 - \text{trace}(\mathbf{\Pi R}) = -\text{trace}(\mathbf{\Pi Q}).$$

The expected number of these in unit time ($t = 1$) is then

$$-\mu \text{trace}(\mathbf{\Pi Q}). \tag{2.6}$$

This is the mutation rate for the process. Care must be taken when comparing two different models in case their underlying mutation rates differ. Given a rate matrix \mathbf{Q} we choose μ such that the overall rate of mutation $-\mu \text{trace}(\mathbf{\Pi Q})$ is one. In this way the length of the branch corresponds to the expected number of mutations per site along that branch, irrespective of the model.

Applying equation (2.6) to the F81 model we obtain a rate of

$$-\mu \text{trace}(\mathbf{\Pi Q}) = \mu(\pi_A(1 - \pi_A) + \pi_C(1 - \pi_C) + \pi_G(1 - \pi_G) + \pi_T(1 - \pi_T)),$$

so, given π_A, \dots, π_T we would set

$$\mu = [\pi_A(1 - \pi_A) + \pi_C(1 - \pi_C) + \pi_G(1 - \pi_G) + \pi_T(1 - \pi_T)]^{-1}$$

to normalize the rates.

2.2.6 Probability of sequence evolution on a tree

We now extend the model for sequence evolution to evolution on a phylogeny. We are still concerned, at this point, with the evolution of a single site. Because of independence between sites, the probability of a set of sequences evolving is just the product of the probabilities for the individual sites.

Each site i in a sequence determines a *character* on the leaves: a function χ_i from the leaf set to the set of states \mathcal{E} . An *extension* $\hat{\chi}_i$ of a character χ_i is an assignment of states to *all* of the nodes in the tree that agrees with χ_i on the leaves.

We define the *probability of an extension* as the probability of the state at the root (given by the stationary distribution) multiplied by the probabilities of all the changes (or conservations) down each branch in the tree. If we use b_{uv} to denote the length of the branch between node u and node v , and let $\hat{\chi}_i(v)$ denote the state assigned to v , then we have a probability

$$\mathbb{P}[\hat{\chi}_i | \theta] = \pi_{\hat{\chi}_i(v_0)} \prod_{\text{branches } \{u, v\}} \mathbf{P}_{\hat{\chi}_i(u)\hat{\chi}_i(v)}(b_{uv}). \quad (2.7)$$

Here, v_0 is the root of the tree. The probability of site i is then the *marginal probability* over all the extensions $\hat{\chi}_i$ of χ_i :

$$\mathbb{P}[\chi_i | \theta] = \sum_{\hat{\chi}_i \text{ extends } \chi_i} \mathbb{P}[\hat{\chi}_i | \theta]. \quad (2.8)$$

The probability of the complete alignment is simply the product of the probabilities of the sites. The next section gives the details of this fundamental calculation.

Equations (2.7) and (2.8) are perhaps better understood if we consider the problem of *simulating* sequences on a tree. To simulate a site in a sequence we first sample a state at the root from the stationary distribution. Then, working down the tree, we sample the state at the end of a branch (furthest from the root) using the value x already sampled at the beginning of the branch, the length of the branch b , and the probabilities in row x of the transition matrix $\mathbf{P}(b)$. The states chosen (eventually) at the leaves then give the character for one site of our simulated sequences. The probability $\mathbb{P}[\chi_i | \theta]$ equals the probability that the character χ_i could have been generated using this simulation method.

2.3 Likelihood calculation: the basic algorithm

Here we describe the basic algorithm for computing the likelihood $L(\theta) = \mathbb{P}[\chi_i | \theta]$ of a site, given a (rooted) tree, branch lengths, and the model of sequence evolution. The likelihood of an alignment is computed by multiplying the likelihoods for each of the n sites

$$L(\theta) = \prod_{i=1}^n \mathbb{P}[\chi_i | \theta]. \quad (2.9)$$

Remember that χ_i is the character (column) corresponding to the i th site in a sequence alignment. Let v be an internal node of the tree, and let $L_i^v(x)$, $x \in \mathcal{E}$ denote the *partial conditional likelihood* defined as:

$$L_i^v(x) = \mathbb{P}[\chi_i^v | \theta, \hat{\chi}_i(v) = x],$$

where χ_i^v is the restriction of the character χ_i to descendants of node v and $\hat{\chi}_i(v)$ is the ancestral state for site i at node v (Fig. 2.2). The value $L_i^v(x)$ is the likelihood at site i for the subtree underlying node v , conditional on state x at v .

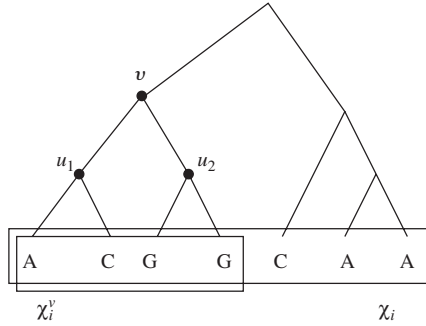


FIG. 2.2. Illustration of a node v , its children u_1, u_2 , the character χ_i and its restriction χ_i^v to the subtree rooted at v .

The likelihood of the complete character χ_i can be expressed as:

$$\mathbb{P}[\chi_i | \theta] = \sum_{x \in \mathcal{E}} \mathbb{P}[\hat{\chi}(v_0) = x] L_i^{v_0}(x), \quad (2.10)$$

where v_0 is the root node. The probability $\mathbb{P}[\hat{\chi}(v_0) = x]$ equals the probability for x under the stationary distribution, π_x .

The function $L_i^v(x)$ satisfies the recurrence

$$L_i^v(x) = \left(\sum_{y \in \mathcal{E}} \mathbf{P}_{xy}(t_1) L_i^{u_1}(y) \right) \left(\sum_{y \in \mathcal{E}} \mathbf{P}_{xy}(t_2) L_i^{u_2}(y) \right), \quad (2.11)$$

for all internal nodes v , where u_1 and u_2 are the children of v and t_1, t_2 are the lengths of the branches connecting them to v . Equation (2.11) results from the independence of the processes in the two subtrees below node v . For leaf l , we have

$$L_i^l(x) = \begin{cases} 1, & \text{if } \chi_i(l) = x, \\ 0, & \text{otherwise.} \end{cases}$$

Note that equation (2.11) can be easily extended to nodes v with more than two children.

The transition probabilities $P_{xy}(t_1)$ and $P_{xy}(t_2)$ are determined from equation (2.2). As observed above, this requires the diagonalization of the rate matrix \mathbf{Q} . However we need only perform this diagonalization once, after which point it only takes $O(c)$ operations, where c is the size of the state set, to evaluate each probability.

The above calculation was defined on a rooted tree. For a time reversible, stationary process, however, the location of the root does not matter: the likelihood value is independent of the position of the root [14]. As well, the logarithm of the likelihood is usually computed rather than the likelihood itself. The product in equation (2.9) becomes a summation if the log-likelihood is computed.

Calculating the log-likelihood of a tree therefore involves

- (i) diagonalization of \mathbf{Q} ;
- (ii) for each branch of the tree, taking the exponential of $\mathbf{Q}ut$, where t is the branch length;
- (iii) for every site and every possible state, applying equation (2.11) using a post-order traversal of the tree;
- (iv) taking the logarithm and summing over sites.

Recall that c is the number of states, m the number of leaves, and n the number of sites. Step (i) can be performed in $O(c^3)$ time using standard numerical techniques. Step (ii) takes $O(mc^3)$ time. Step (iii) takes $O(mnc^2)$ time, and step (iv) takes $O(n)$ time. The whole algorithm therefore takes $O(mc^3 + mnc^2)$ time. Step (iii) is the most computationally expensive step in virtually every application.

2.4 Likelihood calculation: improved models

The calculation presented above applies to standard Markov models of sequence evolution, assuming a single, common process to all sites and in all lineages, and independent sites. Actual molecular evolutionary processes often depart from these assumptions. We now introduce likelihood calculation under more realistic models of sequence evolution, with the aim of improving phylogenetic estimates and of learning more about the evolutionary forces that drive sequence variation.

2.4.1 Choosing the rate matrix

The choice of rate matrix (generator) \mathbf{Q} is an important part of the modelling process. The rate matrix has $c(c-1)$ non-diagonal entries, where c is the number of states. Thus the number of off-diagonal entries equals 12 for DNA sequences, 180 for amino-acid sequences, and 3660 for codon sequences. This number is halved if we also have time reversibility. Allowing one free parameter per rate is not appropriate; one has to introduce constraints in order to reach a reasonable number of free parameters, preferably representing biologically meaningful features of evolutionary processes.

In practice, the features of \mathbf{Q} are determined empirically. For example, in DNA sequences it has been observed that *transitions* (mutations between A and G or between C and T) are more frequent than *transversions* (other mutations). The HKY model [29] incorporates this observation into the rate matrix:

$$\mathbf{Q} = \begin{bmatrix} -(\pi_Y + \kappa\pi_G) & \pi_C & \kappa\pi_G & \pi_T \\ \pi_A & -(\pi_R + \kappa\pi_T) & \pi_G & \kappa\pi_T \\ \kappa\pi_A & \pi_C & -(\pi_Y + \kappa\pi_A) & \pi_T \\ \pi_A & \kappa\pi_C & \pi_G & -(\pi_R + \kappa\pi_C) \end{bmatrix}.$$

As before, $\pi_R = \pi_A + \pi_G$ and $\pi_Y = \pi_C + \pi_T$.

This matrix is the same as that for F81, except for an extra parameter κ affecting the relative rate of mutations within purines or within pyrimidines.

When $\kappa = 1.0$ we obtain the F81 model again. When $\kappa > 1.0$ the rate of transitions is greater than the rate of transversions. A large body of literature discusses the merits of various parameterizations of rate matrices for DNA, protein and codon models (e.g. [49]). We do not review this issue here. The above-described basic likelihood calculation procedure applies whatever the parameterization.

Non-homogeneous models of sequence evolution, in which distinct branches of the tree have distinct rate matrices, have been introduced for modelling variations of the selective regime of protein coding genes [60], or variations of base composition in DNA (RNA) sequences [22]. The calculation of transition probabilities along branches ($\mathbf{P}_{xy}(t)$ in equation (2.11)) should be modified accordingly, using the appropriate rate matrix for each branch. When the distinct rate matrices have unequal equilibrium frequencies [22, 59], the process becomes non-stationary: stationary frequencies are never reached because they vary in time. In this case, the likelihood function becomes dependent on the location of the root, and the ancestral frequency spectrum ($\mathbb{P}[\hat{\chi}(v_0) = x]$ in equation (2.10)) becomes an additional parameter: it can no longer be deduced from the evolutionary model.

2.4.2 Among site rate variation (ASRV)

A strong and unrealistic assumption of the standard model is that sites evolve at the same rate. In real data sets there are typically fast and slowly evolving sites, mostly as a consequence of variable selective pressure. Functionally important sites are conserved during evolution, while unimportant sites are free to vary.

Yang first introduced likelihood calculation incorporating variable rates across sites [55]. He proposed that the variation of evolutionary rates across sites be modelled by a continuous distribution: the rate of a specific site i is not a constant, but a random variable $r(i)$. The likelihood for site i is calculated by integrating over all possible rates:

$$\mathbb{P}[\chi_i | \theta] = \int_0^\infty \mathbb{P}[\chi_i | r(i) = r, \theta] f(r) dr, \quad (2.12)$$

where f is the probability density of the assumed rates distribution, and where $\mathbb{P}[\chi_i | r(i) = r, \theta]$ is the likelihood for character χ_i conditional on rate $r(i) = r$ for this site. The latter term is calculated by applying recurrence (2.11) after multiplying all of the branch lengths in the tree by r . Typically, a Gamma distribution is used for $f(r)$. Its variance and shape are controlled by an additional parameter that can be estimated from the data by the maximum-likelihood method.

The integration in equation (2.12) must be performed numerically, which is time consuming. In practice, this calculation can be completed only for small trees. For this reason, Yang proposed to assume a discrete, rather than continuous, distribution of rates across sites [56]:

$$\mathbb{P}[\chi_i | \theta] = \sum_{j=1}^g \mathbb{P}[\chi_i | r(i) = r_j, \theta] p_j, \quad (2.13)$$

where g is the assumed number of rate classes and p_j the probability of rate class j . Yang [56] uses a discretized Gamma distribution for the probabilities p_j . The complexity of the likelihood calculation under the discrete-Gamma model of rate variation is $O(mc^3g + mnc^2g)$, that is, essentially g times the complexity of the equal-rate calculation. Using ASRV models typically leads to a large increase of log-likelihood, compared to constant-rate models. The extension of this approach to heterogeneous models of site evolution is the subject of Chapter 5, this volume.

Note that sites are not assigned to rate classes in this calculation. Rather, all possible assignments are considered, and the conditional likelihoods averaged. Sites can be assigned to rate classes following likelihood calculation. The *posterior probability* of rate class j for site Y_i can be defined as:

$$\mathbb{P}[\text{site } i \text{ in class } j] = \frac{p_j P[\chi_i | r(i) = r_j, \theta]}{P[\chi_i | \theta]}, \quad (2.14)$$

where the calculation is achieved using the maximum likelihood estimates of parameters (tree, branch lengths, rate matrix, gamma shape). This equation does not account for the uncertainty in the unknown parameters, an approximate procedure called “empirical Bayesian” [61].

2.4.3 Site-specific rate variation

In models of between-site rate variation, the (relative) rate of a site is constant in time: a slow site is slow, and a fast site fast, in every lineage of the tree. In reality, evolutionary rate might, however, vary in time, if the level of constraint applying to a specific site changes. The notion that the evolutionary rate of a site can evolve was first introduced by Fitch [19], and subsequently modelled by Tuffley and Steel [52] and Galtier [21]. This process has been named *covarion* (for COncomitantly VARIable codON [19]), heterotachy, or site-specific rate variation.

Covarion models typically assume a compound process of evolution. The rate of a given site evolves along the tree according to a Markov process defined in the space of rates. Thus the site evolves in the state space according to a Markov process whose local rate is determined by the outcome of the rate process. A site can be fast in some parts of the tree, but slow in other parts. Such processes are called *Markov-modulated Markov processes* or *Cox processes*. The state process is modulated by the rate process.

Existing models use a discrete rate space: a finite number g of Gamma distributed rates are permitted, just like in discretized ASRV models (see above). Let $\mathbf{r} = (r_j)$ be the vector of allowed rates (size g), let $\text{diag}(\mathbf{r})$ be the diagonal matrix with diagonal entries r_j , and \mathbf{G} be the rate matrix of the rate process, indexed by the rate classes. Let \mathbf{Q} be the rate matrix of the state process. The compound process can be seen as a single process taking values in $\{r_j\} \times \mathcal{E}$, a compound space of size $g \cdot c$. The rate matrix, \mathbf{Z} , of this process can be expressed using the Kronecker operand \otimes . If \mathbf{A} is an $m \times m$ matrix and \mathbf{B} is an $n \times n$

matrix then $\mathbf{A} \otimes \mathbf{B}$ is the $mn \times mn$ matrix

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} \mathbf{A}_{11}\mathbf{B} & \dots & \mathbf{A}_{1m}\mathbf{B} \\ \vdots & \ddots & \vdots \\ \mathbf{A}_{m1}\mathbf{B} & \dots & \mathbf{A}_{mm}\mathbf{B} \end{bmatrix}.$$

The rate matrix \mathbf{Z} can then be expressed as

$$\mathbf{Z} = \text{diag}(\mathbf{r}) \otimes \mathbf{Q} + \mathbf{G} \otimes \mathbf{I}_c, \quad (2.15)$$

where \mathbf{I}_c is the $c \times c$ identity matrix [23]. Likelihood calculation under this model is therefore achieved similarly to the standard model, using a rate matrix of size $g \cdot c$. The complexity of the algorithm becomes $O(mc^3g^3 + mnc^2g^2)$.

As an example, consider the basic covarion model of Tuffley and Steel [52]. This model uses only two different rates: ‘‘on’’ ($r_1 = 1$) and ‘‘off’’ ($r_2 = 0$). The switching between rates is controlled by the rate matrix

$$\mathbf{G} = \begin{bmatrix} -s_1 & s_1 \\ s_2 & -s_2 \end{bmatrix}.$$

To apply the covarion approach with the F81 model we plug in the rate matrix \mathbf{Q} from equation (2.3) to give the rate matrix for the compound process of

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Q} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} -s_1\mathbf{I}_c & s_1\mathbf{I}_c \\ s_2\mathbf{I}_c & -s_2\mathbf{I}_c \end{bmatrix} = \begin{bmatrix} * & \pi_C & \pi_G & \pi_T & s_1 & 0 & 0 & 0 \\ \pi_A & * & \pi_G & \pi_T & 0 & s_1 & 0 & 0 \\ \pi_A & \pi_C & * & \pi_T & 0 & 0 & s_1 & 0 \\ \pi_A & \pi_C & \pi_G & * & 0 & 0 & 0 & s_1 \\ s_2 & 0 & 0 & 0 & * & 0 & 0 & 0 \\ 0 & s_2 & 0 & 0 & 0 & * & 0 & 0 \\ 0 & 0 & s_2 & 0 & 0 & 0 & * & 0 \\ 0 & 0 & 0 & s_2 & 0 & 0 & 0 & * \end{bmatrix}.$$

The values along the diagonal are chosen so that the row sums are all zero. The state set for this process is $\{(A, \text{on}), (C, \text{on}), (G, \text{on}), (T, \text{on}), (A, \text{off}), (C, \text{off}), (G, \text{off}), (T, \text{off})\}$.

2.4.4 Correlated evolution between sites

Independence between sites is a fundamental assumption of standard Markov models of sequence evolution, expressed in equation (2.9). The sites of a functional molecule, however, do not evolve independently in the real world: biochemical interactions between sites are required for stabilizing the structure, and achieving the function, of biomolecules.

Pollock *et al.* proposed a model for relaxing the independence assumption [35].

Consider the joint evolutionary process of any two sites of a protein. The space state for the joint process is $\mathcal{E} \times \mathcal{E}$. Under the assumption of independent

sites, the rate matrix for the joint process is constructed from that of the single-site process (assume reversibility):

$$\begin{aligned}\overline{\mathbf{Q}}_{xx',yy'} &= \mathbf{Q}_{xy} = \mathbf{S}_{xy}\pi_y, \\ \overline{\mathbf{Q}}_{xx',xy'} &= \mathbf{Q}_{x'y'} = \mathbf{S}_{x'y'}\pi'_y, \\ \overline{\mathbf{Q}}_{xx',yy'} &= 0,\end{aligned}\tag{2.16}$$

for $x \neq y$ and $x' \neq y'$, where $\overline{\mathbf{Q}}_{xx',yy'}$ is the rate of change from x to y at site 1, and from x' to y' at site 2 (in $\mathcal{E} \times \mathcal{E}$), where \mathbf{Q}_{xy} and π_x are the rate matrix and stationary distribution for the single-site process with state space \mathcal{E} and where $\mathbf{S} = \mathbf{\Pi}^{-1}\mathbf{Q}$ is a symmetric matrix. The joint rate matrix $\overline{\mathbf{Q}}$ has dimension c^2 .

Modelling non-independence between the two sites involves departing from equation (2.16). This is naturally achieved by amending stationary frequencies. It is easy to show that the stationary frequency $\overline{\pi}_{xx'}$ of state $(x, x') \in \mathcal{E}$ is equal to the $\pi_x\pi'_x$ product under the independence assumption. Non-independence can be introduced by rewriting the above equation as:

$$\begin{aligned}\overline{\mathbf{Q}}_{xx',yy'} &= \mathbf{S}_{xy}\overline{\pi}_{yx'}, \\ \overline{\mathbf{Q}}_{xx',xy'} &= \mathbf{S}_{x'y'}\overline{\pi}_{xy'}, \\ \overline{\mathbf{Q}}_{xx',yy'} &= 0,\end{aligned}\tag{2.17}$$

where $\overline{\pi}_{xx'}$'s are free parameters (possibly some function of π_x 's). This formalization accounts for the existence of frequent and infrequent combinations of states between the two sites, perhaps distinct from the product of marginal site-specific frequencies. Pollock *et al.* applied this idea in a simplified, two-state model of protein evolution [35], to be applied to a specific site pair of interest. The same idea was used by Tillier and Collins [51] when they introduced a model dedicated to paired sites in ribosomal RNA. From an algorithmic point of view, accounting for co-evolving site pairs corresponds to a squaring of the state space size c .

Other models aim at representing the fact that two sites have correlated evolutionary rates [17, 57]. Such models are extensions of the ASRV model in which the distribution of site-specific evolutionary rates are not independent among sites. More specifically, these two studies propose a model in which neighbouring sites have correlated rates, introducing an autocorrelation parameter. The idea was extended by Goldman and coworkers when they assumed distinct categories of rate matrices among amino acid sites, and correlated probabilities of the various categories between neighbouring sites [36, 50].

2.5 Optimizing parameters

So far we have not considered what is really the most difficult and limiting aspect of likelihood analysis in phylogenetics: parameter optimization. The problem of finding the maximum likelihood phylogeny combines continuous and discrete optimization. The optimization of branch lengths (and sometimes other

parameters) on a fixed tree is a continuous optimization problem, while the problem of finding the maximum likelihood tree is discrete. Both components are difficult computationally, and computational biologists have not got much past simple heuristics in either case. While these heuristics are proving highly effective, faster and more accurate algorithms are still needed.

2.5.1 *Optimizing continuous parameters*

Given a fixed tree, it is a non-trivial problem to determine the branch lengths giving the maximum likelihood score. On a hundred taxa tree, there are 197 branches, so we are faced with optimizing a 197 dimensional, non-linear, generally non-convex, function. Chor *et al.* [10] have shown that the function can become almost arbitrarily complex. There can be infinitely many local (or global) optima, even when there are only four taxa and two states. Rogers and Swofford [38] observe that multiple optima arise only infrequently in practice. This was not confirmed by our own, preliminary investigations, where we found it relatively easy to generate situations with multiple optima, especially when there was a slight violation of the evolutionary model.

Almost all of the widely used phylogeny programs improve branch lengths iteratively and one at a time. The general approach is to

1. Choose initial branch lengths (here represented as a vector \mathbf{b}).
2. Repeat for each branch k :
 - (a) Find a real number λ_k so that replacing the length b_k of branch k with $b_k + \lambda_k$ gives the largest likelihood.
 - (b) Replace b_k with $b_k + \lambda_k$ and update the partial likelihood computations (see, for example, the updating algorithm of [1]).
3. If λ_k was small for all branches then return the current branch lengths, otherwise go to step 2.

Implementations differ with respect to the one-dimensional optimization technique used to determine λ_k . The technique used most often is Newton's method (also known as the Newton-Raphson method). The intuitive idea behind Newton's method is to use first and second derivatives to approximate the likelihood function (varying along that branch) by a quadratic function. The branch length is adjusted to equal the minimum of this quadratic function, a new quadratic function is fitted, and the procedure repeats until convergence. The search is constrained so as to maintain non-negative branch lengths. PUZZLE, PAUP*, and PHYML use Brent's method for one-dimensional optimization [4], thereby avoiding the need for partial derivatives. This method is similar to Newton's method, but is more robust. PHYLIP uses a numerical approximation to Newton's method.

Two software packages, NHML and PAUP*, differ from the standard approach and implement a multi-dimensional search, so that more than one branch length is changed at a time. A (fiddly) modification of the pruning algorithm of Section 2.3 can be used to compute the gradient vector and Hessian matrix for a particular set of branch lengths in $O(mnc^3)$ and $O(m^2nc^3)$

time respectively. Hence, *multi-dimensional* Newton Rhapsion and quasi-Newton methods can be implemented fairly efficiently (see [25] for an excellent survey of multi-dimensional optimization methods). A combination of full dimensional and single branch optimization is also possible. One complication is the constraint that branch lengths be non-negative. NHML handles this by defaulting to the *simplex method* (see [25]) when one branch length becomes zero.

Surprisingly, there appears to be no published experimental comparison between single branch and multi-dimensional optimization techniques for likelihood. Our preliminary simulation results indicate that the more sophisticated algorithms will occasionally find better optima, but the increased overhead makes the simple, one branch at a time, approach preferable for extensive tree searches.

2.5.2 Searching for the optimal tree

By far the most widely used method for finding maximum likelihood trees is local search. Using one of several possible methods, we construct an initial tree. We then search through the set of all minor modifications of that tree (see Swofford *et al.* [49] for a survey of these modifications). If we find a modified tree with an improved likelihood, we switch to that tree. The process then continues, each time looking for improved modifications, finally stopping when we reach a local optimum. In practice, users will typically constrain some groups of species, searching only through the smaller set of trees for which these groups are monophyletic (i.e. trees containing these groups as clusters).

There are five standard methods for obtaining an initial tree. Refer to Swofford *et al.* [49] for further details.

- **Randomly generated tree** Used to check for multiple local optima.
- **Distance based tree** Compute a distance matrix for the taxa and apply a distance based method such as Neighbor Joining [39] or BioNJ [24].
- **Sequential insertion** Randomly order the taxa. Construct a tree from the first three taxa. Thereafter, insert the taxa one at a time. At each insertion, place the taxon so that the likelihood is maximized. Some implementations perform local searches after each insertion. One advantage of random sequential insertion is that multiple starting trees can be obtained by varying the insertion order.
- **Star decomposition** Start with a tree with all of the taxa and no internal edges. At each step, choose a pair of nodes to combine, continuing until the tree is fully resolved.
- **Approximate likelihood** Perform a tree search using a criterion that is computationally less expensive than likelihood but chooses similar trees.

A typical maximum likelihood search will involve multiple runs of the starting tree and local search combination. As in all optimization problems there is a risk of getting stuck in a local optimum. To avoid this, it is sometimes desirable to occasionally, and randomly, move to trees with lower likelihood scores. This idea has been formalized in search strategies based on simulated annealing [40], as well

as approaches using genetic algorithms [3, 31]. Vinh and von Haeseler [54] have shown recently that deleting and re-inserting taxa can also help avoid getting trapped in local optima. When multiple searches are run in parallel, information can be communicated between the different searches in order to more rapidly locate areas of tree space with higher likelihoods [30].

2.5.3 *Alternative search strategies*

There has been only a small number of likelihood search methods proposed that differ significantly from the local search framework described above. NJML [34] combines a distance-based method (Neighbor Joining) with maximum likelihood. A partially resolved tree (i.e. a tree with some high degree nodes) is obtained by taking the consensus of a number of NJ bootstrap trees. The method then searches for the tree with maximum likelihood among all trees that contain all of the groups in this partially resolved tree, PhyML [28] gains considerable efficiency by not optimizing all branch lengths for every tree examined. Instead, the algorithm combines moves that improve branch lengths and moves that improve the tree. The advantage of this approach is a considerable gain in speed, as well as the potential to avoid being trapped in some local optima.

A quite different strategy is proposed by Friedman *et al.* [20]. They treat a phylogenetic tree as a graph with vertices and edges. One can estimate the expected mutations between any pair of vertices, then rearrange the tree by removing and adding edges between different pairs of vertices. While the approach has not yet gained widespread acceptance, it represents a completely new way to look at likelihood optimization on trees.

The optimization algorithms implemented in the most widely used phylogenetics packages are summarized in Table 2.1.

2.6 Consistency of the likelihood approach

In this section, we focus on the theoretical underpinnings of the likelihood approach. First we consider the question of consistency: if we have sufficiently long sequences, and the sequence evolution model is correct, will we recover the true tree? As we mentioned above, this does not hold for maximum parsimony. It turns out that maximum likelihood is consistent in most cases. As we shall see, to establish consistency we need to verify an identifiability condition, which ensures that we can distinguish two models from infinite length sequences. We also discuss the robustness of the likelihood approach in coping with model mis-specifications.

2.6.1 *Statistical consistency*

Recall that χ_i represents the character corresponding to the i th site observed in the m sequences and assume that the n sites are independent. The vector of parameters θ includes the tree topology, branch lengths and the parameters of the Markov evolution process. The maximum likelihood estimator $\hat{\theta}_n$ maximizes the

TABLE 2.1. Likelihood algorithms implemented in different software packages. The asterisk indicates that the package implements an algorithm, even if it is not the default algorithm used (as is the case, for example, in PAUP*)

Data	Nucleotides			Proteins and nucleotides					
	PAUP* [48]	fastDNAm1 [32]	NHML [22]	PHYLIP [16]	MOLPHY [1]	PAML [58]	Tree-Puzzle [47]	PHYML [28]	IQPNNI [54]
<i>Approach to branch length optimization</i>									
Single branch per iteration	*	*		*	*	*	*	(a)	*
Multiple branches per iteration									
Newton's method			*						
BFGS (see [25])	*								
Brent's multi-dimension algorithm	*								
Simplex method	*								
<i>Algorithm for one-dimension optimization</i>									
Newton's method (or approximation)	*	*		(b)	*	*			
Brent's one-dimension algorithm	*						*	*	
Subdivision algorithm		*							
<i>Algorithm for the initial tree</i>									
Distance method	*							*	
Random tree	*								
Sequential insertion	*	*	*	*		*			
Star decomposition	*	*		*	*				
Approximate likelihood	*			*	*		*		*
<i>Hill climbing</i>	*	*		*				*	*

Data: which kind of sequence data is analysed. Approach to branch length optimization: whether branches are optimized individually or all at once, and which method is used. Algorithm for one-dimension optimization: which algorithm is used for optimizing a single branch or, in the case of multidimensional optimization, which line search algorithm is used (see [25] for more on line search methods). Algorithm for initial tree: the method used to select the tree (or initial tree when searching). Hill climbing: implements local search that uses the likelihood optimization criterion.

Notes: (a) PHYML combines branch optimization with tree optimization. (b) PHYLIP uses a numerical approximation for first and second derivatives in Newton's method.

likelihood $L(\theta) = \prod_{i=1}^n Pr(\chi_i | \theta)$, or equivalently the normalized log-likelihood

$$l_n(\theta) = \frac{1}{n} \sum_{i=1}^n \log Pr(\chi_i | \theta).$$

If the estimator $\hat{\theta}_n$ is used to estimate the true parameter θ_0 , then it is certainly desirable that the sequence $\hat{\theta}_n$ converges in probability to θ_0 as n tends to ∞ . If this is true, we say that $\hat{\theta}_n$ is *statistically consistent* for estimating θ_0 .

Clearly, the “asymptotic value” of $\hat{\theta}_n$ depends on the asymptotic behaviour of the random functions l_n . There typically exists a deterministic function $l(\theta)$ such that, by the law of large numbers,

$$l_n(\theta) \xrightarrow{\mathbf{P}} l(\theta), \quad \text{for every } \theta.$$

What is expected is that the maximizer $\hat{\theta}_n$ of l_n converges to a unique point θ_0 which, moreover, is the maximum of the function l . This requires two conditions:

(1) *Model identifiability*

A model is said to be *identifiable* if the probability of the observations is not the same under two different values of the parameter:

$$l(\theta) \neq l(\theta_0) \quad \text{for } \theta \neq \theta_0.$$

Identifiability is a natural and a necessary condition: If the parameter is not identifiable then consistent estimators cannot exist.

(2) *Convergence of the likelihood function*

Consistency requires an appropriate form of the functional convergence of l_n to l to ensure the convergence of the maximum of l_n to a maximum of l . There are several situations under which this always holds. The “classical” approach of Wald relies on a continuity argument and a suitable compactification of the parameter set [53]. In the phylogenetic context, Wald’s conditions can be adapted for binary trees [7, 37]. In particular, the continuity of the likelihood reconstruction, with respect to the topology parameter, relies on an argument of Buneman [6].

In a variety of situations of parametric statistical inference, identifiability is trivially fulfilled or it implies restrictive but natural conditions on the parameter space. For most models in the phylogenetic setting, identifiability considerations are the principal difficulty in establishing the consistency of maximum likelihood. As long as the model is identifiable, maximum likelihood estimators are typically consistent.

Note, however, that consistency guarantees identification of the correct parameter values (e.g. the tree topology) with infinite length sequences. In real data situations, the sequence length is finite and no method can be sure to recover the correct parameter values.

2.6.2 Identifiability of the phylogenetic models

In earlier sections, we assumed that there was the same evolutionary model for each branch of the tree. We generalize this here by assigning a different rate matrix $\mathbf{Q}^{(b)}$ to each branch b . The evolutionary *scenario* then comprises the tree topology and the Markov transition matrices across the branches,

$$\mathbf{P}^{(b)}(t) = \exp(\mathbf{Q}^{(b)}t^{(b)}),$$

where $t^{(b)}$ is the length of branch b . Let π^v be the marginal distribution of a site at node v , $\pi^v(x) = \mathbb{P}[\hat{\chi}_i(v) = x]$.

Identifiability requires that two different scenarios (differing in topology or transition matrices or both) cannot induce the same joint distribution of the sites with infinite sequences; if two scenarios were indistinguishable from infinite sequences, there will be no hope that they could be distinguished from observed finite sequences and that maximum likelihood could consistently recover the correct scenario. Here we review what is and what is not known about the identifiability of Markov evolutionary scenarios.

Identical evolution of the sites. Suppose that each site evolves according to the same Markov process, that is, the characters χ_i are independent and identically distributed. Conditions under which identifiability of the full scenario (topology and transition matrices) holds were first established formally by Chang [7].

Identifiability of the topology

Assumption (H): There is a node v with $\pi^v(x) > 0$ for all x , and $\det(\mathbf{P}^{(b)}) \notin \{-1, 0, 1\}$ for all branches b .

Under assumption (H), the topology is identifiable from the joint distribution of the pairs of sites. Assumption (H) is a mild condition which ensures that transition matrices are invertible and not equal to permutation matrices. It enables us to construct an additive tree distance from the character distribution. The so-called *LogDet transform* is a good distance candidate and the tree can be recovered using distance-based methods like those reviewed in Chapter 1, this volume. Identifiability just of the tree was proved by Chang and Hartigan [9] and Steel *et al.* [45] and is more thoroughly discussed in Semple and Steel [42].

Identifiability of the transition matrices Chang showed that we cannot just consider pairwise comparisons of sequences to reconstruct the transition matrices, and that the distribution of triples of sites is required to ensure the identifiability of the full scenario. More precisely, under assumption (H), if moreover the underlying evolutionary tree is binary and the transition matrices belong to a class of matrices that is *reconstructible from rows*, then all of the transition matrices are uniquely determined by the distribution of the triples of sites. Chang's additional condition is somewhat technical: a class of matrices is reconstructible from rows if no two matrices in the class differ only by a permutation of rows. An example of such a class is that in which the diagonal element is always the largest element in a row.

The situation is greatly simplified under the assumptions that the evolution process is stationary and reversible with equilibrium measure π . In this restricted class of Markov models, the distribution of the pairs of sites is enough to determine the full scenario:

Under assumption (H), if the rate matrix is identical on all branches $\mathbf{Q}^{(b)} \equiv \mathbf{Q}$, if it is reversible and the node distribution is the stationary distribution $\pi^v \equiv \pi$, then the (unrooted) topology and the transition matrix is identifiable from the pairwise comparisons of sequences.

In summary, the parameters are identifiable (and hence ML is consistent) not only for the basic models described above, but for far more general scenarios of sequence evolution.

Sites evolving according to different processes. Models that allow different sites to evolve at different rates can be seen as mixtures of Markov models (see Chapter 5, this volume). The difficulty with such heterogeneous models is that a mixture of Markov models is generally not a Markov model and the existence of an additive distance measure to reconstruct a topology, heavily relies upon the Markov property. Baake [2] established that if a rate factor varies from site to site, different topologies may produce identical pairwise distributions. Consequently, identifiability of the topology is lost on the basis of pairwise distributions, even if the distribution of rate factors is known. However, the maximum likelihood method makes use of the full joint distribution of the sites; it can still be expected that conditions of identifiability may be recovered from the complete information of infinite sequences in general heterogeneous models. Nothing has been proved in the general context yet.

Identifiability issues have been discussed under the stationary and reversible assumption. Results have been established by Chang [8], Steel *et al.* [46] and are summarized in Semple and Steel [42].

Suppose that the Markov process is stationary and time reversible, and that on every branch b , all sites evolve according to the same rate matrix \mathbf{Q} multiplied by a rate factor r selected according to a probability distribution $f(r)$. The transition matrix for the sites evolving at rate factor r is

$$\mathbf{P}(b) = \exp\left(r\mathbf{Q}t^{(b)}\right), \quad r \text{ drawn with distribution } f.$$

Under assumption (H), the topology and the full scenario are identifiable if

- f is completely specified up to one or several free parameters, or
- f is unknown but a molecular clock applies, that is, all of the leaves of the tree are equidistant from the root.

The case with f completely specified is formally identical to the situation with constant rates, if the LogDet transform is replaced by an appropriate tree distance based on the moment-generating function of f . One tractable case is where f is a Gamma distribution and its density function is governed by one

parameter estimated from the data (see Section 2.4.2). Without a parameterized form of the distribution f or without strong assumptions such as a molecular clock, different choices of f and the transition matrices may be combined with any tree to produce the same joint distribution of the sites.

Tuffley and Steel [52] analysed a simple covarion model and compared it with the rates-across-sites model of Yang [55]. They showed that the two models cannot be distinguished from the pairwise distribution of the sites but argued that the two models could indeed be identified from the full joint distribution, provided the number of leaves is at least four. A proof of the identifiability of Site-specific Rate Variation models (see Section 2.4.3) remains to be done. However, these models are already implemented [21] and experience indicates that they should be identifiable.

2.6.3 *Coping with errors in the model*

Current implementations are restricted to stationary and reversible models: homogeneous or ASRV models, including mixed invariable sites and gamma-distributed rates. In these cases, the models are identifiable under mild conditions, and maximum likelihood will consistently estimate the tree topology, the branch lengths and the parameters of the Markov evolution process.

Several authors have published examples where maximum likelihood does not recover the true tree [8, 15]. However, none of these constitute a counterexample to the consistency of maximum likelihood methods since, in each case, the basic conditions for consistency are not fulfilled. They either lack identifiability, or the true model is not a member of the class of models considered.

We have stressed several times that the models used in likelihood analysis are simplifications of the actual processes. For this reason, it is essential that we consider the effect of *model misspecification*. Suppose we postulate a model $\{\mathcal{P}_\theta, \theta \in \Theta\}$; however, the model is misspecified in that the true distribution P that generated the data does not belong to the model. For instance, we can perform a maximum likelihood reconstruction with a single stationary Markov model whereas the observations were truly generated by a mixture of Markov models (Chapter 5, this volume). If we use the postulated model anyway, we obtain an estimate $\hat{\theta}_n$ from maximizing the likelihood. What is the asymptotic behaviour of $\hat{\theta}_n$?

Under conditions (1) and (2) (Section 2.6.1), we can prove that $\hat{\theta}_n$ converges to the value θ_0 that maximizes the function $\theta \rightarrow l(\theta)$. The model \mathcal{P}_{θ_0} can be viewed as the “projection” of the true underlying distribution P on the family $\{\mathcal{P}_\theta\}$ using the so-called *Kullback–Leibler divergence* as a distance measure. If the model \mathcal{P}_{θ_0} is not too far off from the truth, we can hope that the estimator $\mathcal{P}_{\hat{\theta}}$ is a reasonable approximation for the true model P . At least, this is what happens in standard classical models, which are nicely parametrized by Euclidean parameters [53].

In the phylogenetic setting, things are complicated by the presence of a discrete non-Euclidean tree parameter. The standard theory does not extend in a straightforward manner. It is not surprising that the above-cited

“counterexamples” all display tree topologies where long branches are separated by short branches; these situations typically favour a lack of robustness. To what extent can likelihood reconstructions recover the true topology when the evolution model is misspecified? A better understanding of the uncertainty in tree estimation is an important direction for future work, so that we can quantify the robustness of likelihood methods and improve testing procedures (see Chapter 4, this volume).

2.7 Likelihood ratio tests

Once a model is developed and the likelihood is optimized, that model may be used to carry out many different statistical tests. In traditional hypothesis testing one often chooses a *null hypothesis* H_0 defined as the absence of some effect; this can be viewed as testing whether some parameter values are equal to zero. For example, testing whether the proportion of invariant sites is zero, or whether there is no rate heterogeneity between sites. If the increase in log-likelihood from raising the proportion of invariant sites from its value under H_0 , that is, 0, to its maximum likelihood estimation is “significant” in some sense, then H_0 is rejected at level α (where α is the probability of rejecting H_0 when it is indeed true). Otherwise, we say that the data at hand do not allow us to reject H_0 ; the proportion of invariant sites may indeed be positive, but we cannot detect this.

Suppose that H_0 is derived from a full alternative H_1 by setting certain parameter values to 0. We can then define sets Θ_0 and Θ_1 such that H_0 corresponds to the situation that the true parameter θ is in $\Theta_0 \subseteq \Theta_1$, and H_1 corresponds to the case $\theta \in \Theta_1 - \Theta_0$. A natural testing idea is to compare the values of the log-likelihood computed under H_0 and H_1 , respectively. The corresponding normalized test statistic is called the (log)*likelihood ratio statistic*.

$$\text{LR} = -2 \left[\max_{\theta \in \Theta_0} \log(L(\theta)) - \max_{\theta \in \Theta_1} \log(L(\theta)) \right].$$

The statistic LR is asymptotically chi-squared distributed under the null hypothesis. The decision rule becomes: reject H_0 if the value of the likelihood ratio statistic exceeds the upper α -quantile of the chi-square distribution. Likelihood ratio tests turn out to be the most powerful tests in an asymptotic sense and in special cases. Thus they are widely used as byproducts of maximum likelihood estimation. However, it is important to realize that their validity heavily relies on two main conditions: H_0 is a simpler model *nested* within the full model H_1 and the correct model belongs to the full model H_1 . For example, in testing whether the proportion of invariant sites is zero, the latter condition implies that the estimated topology is correct and the true rate distribution belongs to gamma + invariant distributions.

Several papers have recently documented the incorrect use and interpretation of standard tests in phylogenetics, due to improper specifications of the test hypotheses [26], or to biases in the asymptotic test distributions [33] or to model misspecification [5]. Ewens and Grant [12] present examples where an

inappropriate use of the LR statistic can cause problems. We review here the assumptions that have to be fulfilled to ensure the validity of likelihood ratio tests and we make precise some restrictions on their applicability. In particular, tests comparing tree topologies cannot use directly the asymptotic framework of likelihood ratio testing.

2.7.1 When to use the asymptotic χ^2 distribution

Suppose a sequence of maximum likelihood estimators $\hat{\theta}_n$ is consistent for a parameter θ that ranges over an open subset of R^p . This is typically true under Wald's conditions and identifiability (see Section 2.6). The next question of interest concerns the order at which the discrepancy $\hat{\theta}_n - \theta$ converges to zero. A standard result says that the sampling distribution of the maximum likelihood estimator has a limiting normal distribution

$$\sqrt{n}(\hat{\theta}_n - \theta) \rightarrow \mathcal{N}(0, \mathbf{i}^{-1}(\theta)), \quad \text{as } n \rightarrow \infty,$$

where $\mathbf{i}(\theta)$ is the Fisher *information matrix*, that is, the $p \times p$ matrix whose elements are the negative of the expectation of all second partial derivatives of $\log L(\theta)$. The convergence in distribution means roughly that $(\hat{\theta}_n - \theta)$ is $\mathcal{N}(0, (n\mathbf{i}(\theta))^{-1})$ -distributed for large n . It implies that the maximum likelihood estimator is asymptotically of minimum variance and unbiased, and in this sense optimal [53].

Suppose we wish to test the null hypothesis H_0 that is nested in the full parameter set of the model of the analysis, say H_1 . Write $\hat{\theta}_{n,0}$ and $\hat{\theta}_n$ for the maximum likelihood estimators of θ under H_0 and H_1 , respectively. The likelihood ratio test statistic is

$$\text{LR} = -2 \left[\log L(\hat{\theta}_{n,0}) - \log L(\hat{\theta}_n) \right].$$

If both H_0 and H_1 are "regular" parametric models that contains θ as an inner point, then, both $\hat{\theta}_{n,0}$ and $\hat{\theta}_n$ can be expected to be asymptotically normal with mean θ and we obtain the approximation under H_0

$$\text{LR} \sim \sqrt{n}(\hat{\theta}_n - \hat{\theta}_{n,0})^t \mathbf{i}(\theta) \sqrt{n}(\hat{\theta}_n - \hat{\theta}_{n,0}).$$

Then the likelihood ratio statistic can be shown to be asymptotically distributed as a quadratic form in normal variables. The law of this quadratic form is a chi-square distribution with $p - q$ degrees of freedom, where p and q are the dimensions of the full and null hypotheses.

The main conditions for this theory to apply are that the null and full hypothesis H_0 and H_1 are equal to R^q and R^p (or are locally identical to those linear spaces), and that the maximum likelihood estimator finds a non-boundary point where the likelihood function is differentiable.

2.7.2 Testing a subset of real parameters

The requirement that the parameters of interest be real numbers is not met if the tree topology is estimated as part of the maximizing procedure. Thus for the moment we assume that the tree topology is given. θ represents here the

scalar parameters, that is, the branch lengths and/or parameters of the evolution process.

Suppose that we wish to test a general linear hypothesis $H_0: A\theta = 0$, where A is a contrast matrix of rank k (i.e. there are $p - k$ free parameters to estimate under H_0). For example, $A\theta = 0$ could correspond to the situation where a particular parameter is zero, in which case $k = 1$. For large n , it can be assumed in this case that LR has a chi-square distribution with k degrees of freedom under H_0 . LR is typically computed by examining successively more complex models, for example, to test whether increasing the number of parameters of the rate matrix \mathbf{Q} yields a significant improvement in model fitting, with respect to the chosen topology.

The LR test is based on the assumption that the tree topology and the evolutionary model are correct. If it is not the case, the induced model bias can make tests reject H_0 too often, or too rarely [5]. In practice, phylogenetic models are always misspecified to a degree. This means that one has to be cautious in interpreting test results for any real data, even if the test is well-founded with respect to theory.

2.7.3 Testing parameters with boundary conditions

We have assumed that the topology is given; even under this restriction, the chi-square approximation fails in a number of simple examples. The “local linearity” of the hypotheses H_0 and H_1 mentioned above is essential for the chi-square approximation. If H_0 defines a region in the parameter space where some parameters are not specified, there is no guarantee in general that the distribution of the test statistic is the same for all points in this region. In tests of one-sided hypotheses, the null hypothesis is no longer locally linear at its boundary points. In this case, however, the testing procedure can be adapted: the asymptotic null distribution of the LR statistic is not chi-squared, but the distribution of a certain functional of a Gaussian vector [41].

A related example arises when some parameters of interest lie on the boundary of the parameter space Θ_1 . Usual boundary conditions are that the branch lengths, the proportion of invariant sites or the shape of a gamma distribution of site substitution rates have non-negative values and difficulties occur when testing whether those parameters are zero. Boundary related problems can also affect tests of the molecular clock. Ota *et al.* [33] derived the appropriate corrections to the asymptotic distributions of the likelihood ratio test statistics, which turn out to be a mixed combination of chi-square distributions and the Dirac function at 0.

2.7.4 Testing trees

When the tree topology is estimated as part of the testing procedure, the conditions derived at the end of Section 2.7.1 are not fulfilled. This is essentially because the tree topology is not a real parameter. Moreover, phylogenetic models displaying different tree topologies are in general not nested. For all these reasons,

tests involving estimated topologies are simply outside the scope of the likelihood ratio tests theory.

Tests involving topologies are thoroughly discussed in Chapter 4, this volume, and alternatives to the classical LR testing procedure are proposed. Another promising testing framework is provided by the likelihood-based tests of multiple tree selection developed in the papers by Shimodaira *et al.* [43, 44]. The model selection approach aims at testing which model is better than the other, while the object of the likelihood ratio test is to find out the correct model. This offers a more flexible approach to model testing, where different topologies combined with different evolution processes can be compared.

2.8 Concluding remarks

Molecular phylogeny is a stimulating topic that lies at the boundary of biology, algorithmics, and statistics, as illustrated in this chapter. The three domains have progressed considerably during the last twenty years: data sets are much bigger, models much better, and programs much faster. Some problems, however, still have to be solved. Not every model that we would want to use permits feasible likelihood calculation. Models for partially relaxing the molecular clock, for example, are highly desirable but currently not tractable in the ML framework. As far as algorithmics is concerned, we have already stressed the probable non-optimality of the optimization algorithms used in the field, a problem worsened by the fact that not all algorithms are published. The statistics of phylogeny also require some clarification, as illustrated in Sections 2.6 and 2.7. The problem of model choice, for example (which model to choose for a given data set), is not really addressed in a satisfactory way in current literature.

An important issue, finally, is the problem of combining data from different genes (the supertree problem). Most approaches to this question have come from combinatorics, while a statistical point of view should be the appropriate one. This would require research into the parametrization of the multi-gene model, and the ability of ML methods to cope with missing data. Recent progress in this area is surveyed in Chapter 5, this volume.

Acknowledgements

We thank Olivier Gascuel and two referees for helpful comments on an earlier version of this chapter. Thanks also to Rachel Bevan, Trevor Bruen, Olivier Gauthier and Miguel Jette for helping with proof-reading. N. G. and M.-A. P. were supported by ACI NIM, ACI IMPBIO, and EPML 64 (CNRS-STIC).

References

- [1] Adachi, J. and Hasegawa, M. (1996). MOLPHY 2.3, programs for molecular phylogenetics based on maximum likelihood. Research Report, Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo.

- [2] Baake, E. (1998). What can and what cannot be inferred from pairwise sequence comparisons? *Mathematical Biosciences*, **154**, 1–22.
- [3] Brauer, M., Holder, M., Dries, L., Zwickli, D., Lewis, P., and Hillis, D. (2002). Genetic algorithms and parallel processing in maximum likelihood phylogeny inference. *Molecular Biology and Evolution*, **19**, 1717–1726.
- [4] Brent, R. (1973). *Algorithms for Minimization without Derivatives*. Prentice-Hall, Englewood Cliffs, NJ.
- [5] Buckley, T.R. (2002). Model misspecification and probabilistic tests of topology: Evidence from empirical data sets. *Systematic Biology*, **51**(3), 509–523.
- [6] Buneman, P. (1971). The recovery of trees from measures of dissimilarity. In *Mathematics in the Archaeological and Historical Sciences* (ed. F. Hodson, D. Kendall, and P. Tautu), pp. 387–395. Edinburgh University Press, Edinburgh.
- [7] Chang, J.T. (1996). Full reconstruction of Markov models on evolutionary trees: Identifiability and consistency. *Mathematical Biosciences*, **137**, 51–73.
- [8] Chang, J.T. (1996). Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. *Mathematical Biosciences*, **134**, 189–215.
- [9] Chang, J.T. and Hartigan, J.A. (1991). Reconstruction of evolutionary trees from pairwise distributions on current species. In *Computing Science and Statistics: Proceeding of the 23rd Symposium on the Interface* (ed. E.M. Keramidas), pp. 254–257. Interface Foundation, Fairfax Station, VA.
- [10] Chor, B., Holland, B.R., Penny, D., and Hendy, M. (2000). Multiple maxima of likelihood in phylogenetic trees: An analytic approach. *Molecular Biology and Evolution*, **17**, 1529–1541.
- [11] Edwards, A.W.F. (1972). *Likelihood*. Cambridge University Press, Cambridge.
- [12] Ewens, W.J. and Grant, G.R. (2001). *Statistical Methods in Bioinformatics*. Springer-Verlag, New York.
- [13] Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, **27**, 401–410.
- [14] Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, **17**, 368–376.
- [15] Felsenstein, J. (2003). *Inferring Phylogenies*. Sinauer Associates Inc., MA.
- [16] Felsenstein, J. (2004). PHYLIP 3.6: The phylogeny inference package.
- [17] Felsenstein, J. and Churchill, G.A. (1996). A hidden Markov model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution*, **13**(1), 93–104.
- [18] Fisher, R.A. (1922). The mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A*, **222**, 309–368.

- [19] Fitch, W.M. (1971). Rate of change of concomitantly variable codons. *Journal of Molecular Evolution*, **1**(1), 84–96.
- [20] Friedman, N., Ninio, M., Pe'er, I., and Pupko, T. (2002). A structural EM algorithm for phylogenetic inference. *Journal of Computational Biology*, **9**, 331–353.
- [21] Galtier, N. (2001). Maximum-likelihood phylogenetic analysis under a covarion-like model. *Molecular Biology and Evolution*, **18**(5), 866–873.
- [22] Galtier, N. and Gouy, M. (1998). Inferring pattern and process: Maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Molecular Biology and Evolution*, **15**(7), 871–879.
- [23] Galtier, N. and Jean-Marie, A. (2004). Markov-modulated Markov chains and the covarion process of molecular evolution. *Journal of Computational Biology*, in press, **11**(4), 727–733.
- [24] Gascuel, O. (1997). BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, **14**(7), 685–695.
- [25] Gill, P., Murray, W., and Wright, M. (1982). *Practical Optimization*. Academic Press, London-New York.
- [26] Goldman, N., Anderson, J.P., and Rodrigo, A.G. (2000). Likelihood-based tests of topologies in phylogenetics. *Systematic Biology*, **49**(4), 652–670.
- [27] Golub, G.H. and van Loan, C.F. (1996). *Matrix Computations* (3rd edn). John Hopkins University Press, Baltimore, MD.
- [28] Guindon, S. and Gascuel, O. (2003). A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, **52**(5), 696–704.
- [29] Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating the human-ape split by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, **22**, 160–174.
- [30] Lemmon, A. and Milinkovitch, M. (2002). The metapopulation genetic algorithm: An efficient solution for the problem of large phylogeny estimation. *Proceedings of National Academy of Science USA*, **99**, 10516–10521.
- [31] Lewis, P. (1998). A genetic algorithm for maximum likelihood phylogeny inference using nucleotide sequence data. *Molecular Biology and Evolution*, **15**, 277–283.
- [32] Olsen, G., Matsuda, H., Hagsstrom, R., and Overbeek, R. (1994). fastDNAm1: A tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Computational Applications in Biosciences*, **10**, 41–48.
- [33] Ota, R., Waddell, P.J., Hasegawa, M., Shimodaira, H., and Kishino, H. (2000). Appropriate likelihood ratio tests and marginal distributions for evolutionary tree models with constraints on parameters. *Molecular Biology and Evolution*, **17**(5), 652–670.

- [34] Ota, S. and Li, W.H. (2000). NJML: A hybrid algorithm for the neighbor-joining and maximum likelihood methods. *Molecular Biology and Evolution*, **17**(9), 1401–1409.
- [35] Pollock, D.D., Taylor, W.R., and Goldman, N. (1999). Coevolving protein residues: Maximum likelihood identification and relationship to structure. *Journal of Molecular Biology*, **287**(1), 187–198.
- [36] Robinson, D.M., Jones, D.T., Kishino, H., Goldman, N., and Thorne, J.L. (2003). Protein evolution with dependence among codons due to tertiary structure. *Molecular Biology and Evolution*, **20**, 1692–1704.
- [37] Rogers, J.S. (1997). On the consistency of maximum likelihood estimation of phylogenetic trees from nucleotide sequences. *Systematic Biology*, **46**, 1079–1085.
- [38] Rogers, J.S. and Swofford, D. (1999). Multiple local maxima for likelihoods of phylogenetic trees: A simulation study. *Molecular Biology and Evolution*, **16**, 1079–1085.
- [39] Saitou, N. and Nei, M. (1987). The neighbor-joining method: A new method for reconstruction of phylogenetic trees. *Molecular Biology and Evolution*, **4**, 406–425.
- [40] Salter, L. and Pearl, D. (2001). Stochastic search strategy for estimation of maximum likelihood phylogenetic trees. *Systematic Biology*, **50**, 7–17.
- [41] Self, S.G. and Liang, K. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, **82**, 605–610.
- [42] Semple, C. and Steel, M. (2003). *Phylogenetics*. Oxford University Press, Oxford(!).
- [43] Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. *Systematic Biology*, **51**, 492–508.
- [44] Shimodaira, H. and Hasegawa, M. (1999). Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution*, **16**, 1114–1116.
- [45] Steel, M., Hendy, M.D., and Penny, D. (1998). Reconstructing probabilities from nucleotide pattern probabilities: A survey and some new results. *Discrete Applied Mathematics*, **88**, 367–396.
- [46] Steel, M., Szekely, L.A., and Hendy, M.D. (1994). Reconstructing trees when sequence sites evolve at variable rates. *Journal of Computational Biology*, **1**, 153–163.
- [47] Strimmer, K. and von Haeseler, A. (1996). Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies. *Molecular Biology and Evolution*, **13**, 964–969.
- [48] Swofford, D. (1998). *PAUP**. *Phylogenetic Analysis Using Parsimony (*and other Methods)*. Version 4. Sinauer Associates, Sunderland, MA.

- [49] Swofford, D., Olsen, G.J., Waddell, P.J., and Hillis, D.M. (1996). Phylogenetic inference. In *Molecular Systematics* (2nd edn) (ed. D. Hillis, C. Moritz, and B. Mable), pp. 438–514. Sinauer, Sutherland, MA.
- [50] Thorne, J.L., Goldman, N., and Jones, D.T. (1996). Combining protein evolution and secondary structure. *Molecular Biology and Evolution*, **13**(5), 666–673.
- [51] Tillier, E.R.M. and Collins, R.A. (1998). High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal RNA. *Genetics*, **148**, 1993–2002.
- [52] Tuffley, C. and Steel, M.A. (1998). Modeling the covarion hypothesis of nucleotide substitution. *Mathematical Biosciences*, **147**, 63–91.
- [53] Van der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge University Press.
- [54] Vinh, L.S. and von Haeseler, A. (2004). IQPNNI: Moving fast through tree space and stopping in time. *Molecular Biology and Evolution*, **21**, 1565–1571.
- [55] Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution*, **10**(6), 1396–1401.
- [56] Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution*, **39**(3), 306–314.
- [57] Yang, Z. (1995). A space-time process model for the evolution of DNA sequences. *Genetics*, **139**, 993–1005.
- [58] Yang, Z. (2000). Phylogenetic analysis by maximum likelihood (PAML), version 3.0.
- [59] Yang, Z. and Roberts, D. (1995). On the use of nucleic acid sequences to infer early branchings in the tree of life. *Molecular Biology and Evolution*, **12**(3), 451–458.
- [60] Yang, Z., Swanson, W.J., and Vacquier, V.D. (2000). Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. *Molecular Biology and Evolution*, **17**(10), 1446–1455.
- [61] Yang, Z. and Wang, T. (1995). Mixed model analysis of DNA sequence evolution. *Biometrics*, **51**(2), 552–561.