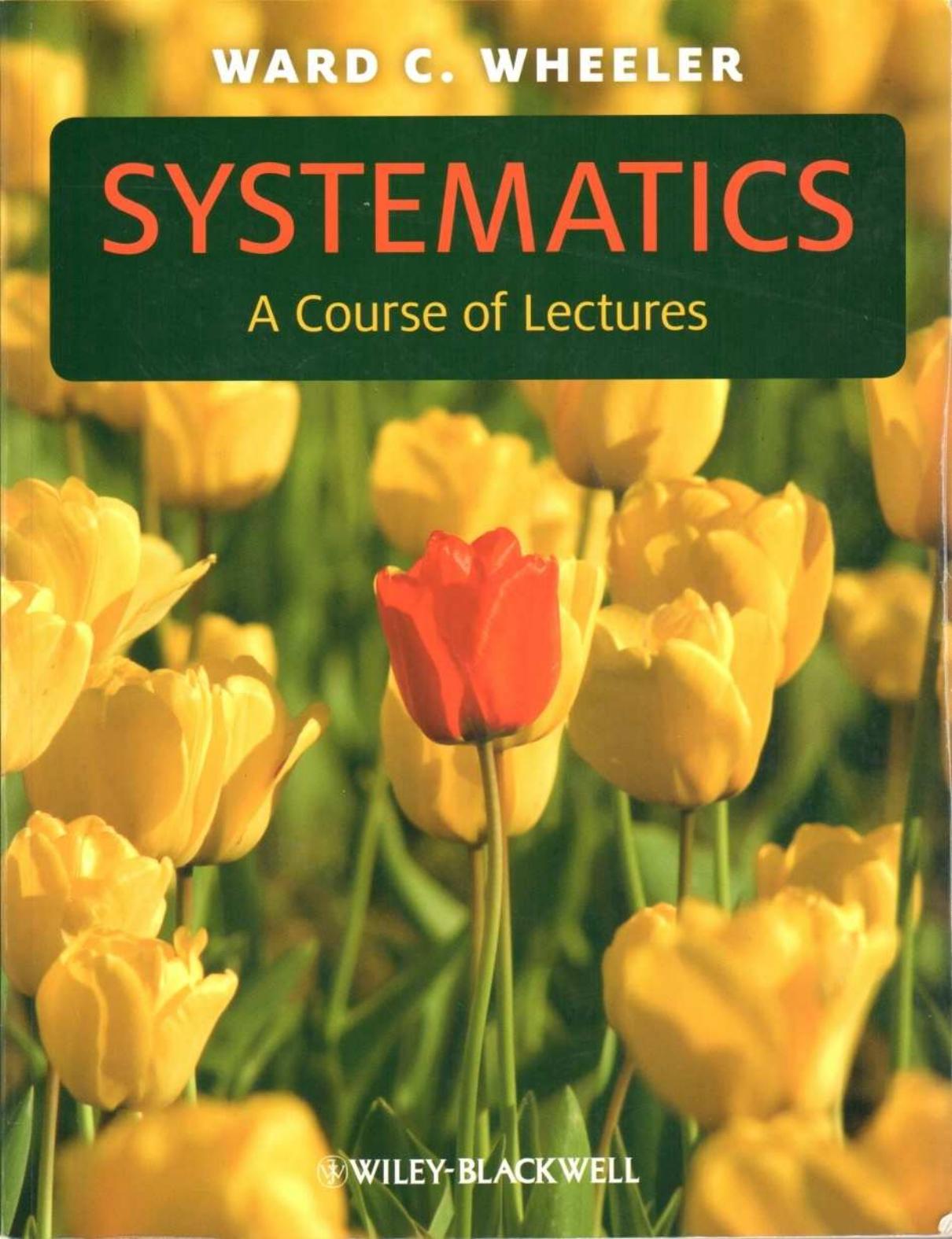
WARD C. WHEELER

SYSTEMATICS

A Course of Lectures



WILEY-BLACKWELL

Systematics

Systematics: A Course of Lectures

Ward C. Wheeler

WILEY-BLACKWELL
A John Wiley & Sons, Ltd., Publication

This edition first published 2012 © 2012 by Ward C. Wheeler

Wiley-Blackwell is an imprint of John Wiley & Sons, formed by the merger of Wiley's global Scientific, Technical and Medical business with Blackwell Publishing.

Registered office: John Wiley & Sons, Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial offices: 9600 Garsington Road, Oxford, OX4 2DQ, UK
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK
111 River Street, Hoboken, NJ 07030-5774, USA

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com/wiley-blackwell.

Library of Congress Cataloging-in-Publication Data has been applied for
9780470671702 (hardback)
9780470671696 (paperback)

A catalogue record for this book is available from the British Library.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Set in Computer Modern 10/12pt by Laserwords Private Limited, Chennai, India

1 2012

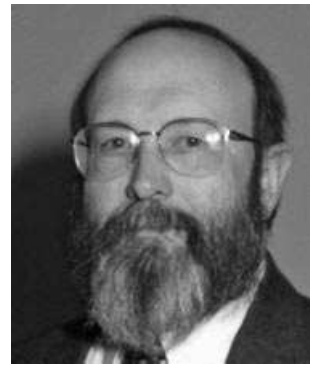
Chapter 11

Optimality Criteria—Likelihood

The previous two chapters discussed methods to determine the cost of a tree based on overall distance and the minimization of weighted transformations. We discuss here the determination of tree cost using stochastic models of character change optimizing the probability of the observed data on T given some set of parameters. This probability is proportional to the likelihood function of Section 6.1.7 and is referred to as the *maximum likelihood* (ML) criterion.

As with parsimony, ML methods assign median (ancestral) states (either in an optimal or average context) such that the overall likelihood of the tree is maximized. Unlike minimization-based parsimony, ML methods require explicit models of character evolution (as opposed to edit cost regimes) and edge parameters (branch lengths; parsimony requires none) to determine tree optimality.

The presentation here will also divide characters into static and dynamic types since they require different analytical techniques.



Joseph Felsenstein

11.1 Motivation

One might explore alternate optimization criteria for their own sake. ML, however, was proposed in the context of purported problems with parsimony analysis. Although Camin and Sokal (1965) and Farris (1973a) had discussed ML methods, Felsenstein (1973) was the first to identify concerns with parsimony and advocate ML as a solution. Much of the discussion centering on the relative merits of parsimony and likelihood in systematics is based on the simple scenario described by Felsenstein (1978).

11.1.1 Felsenstein's Example

Felsenstein posited a four-taxon example (Fig. 11.1) with a simple model of change in binary characters to make his point. In this scenario, there are taxa

Systematics: A Course of Lectures, First Edition. Ward C. Wheeler.

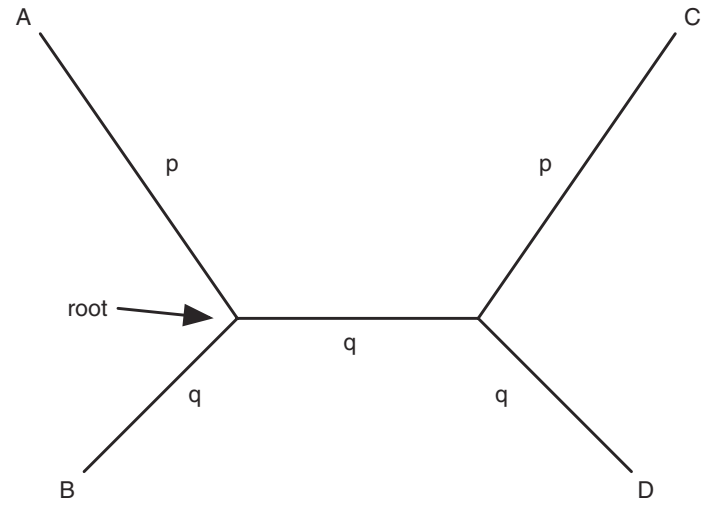


Figure 11.1: Felsenstein (1978) scenario for the statistical inconsistency of parsimony. Probability calculations begin at the root vertex.

A, B, C, and D related by a tree. A and B are on one side of a central split and C and D the other. All characters are posited to have state 0 at the arbitrarily labelled root position. The probabilities of change on each branch are either p or q as labelled. The probabilities of character change are symmetrical so that for all characters $pr(0 \rightarrow 1) = pr(1 \rightarrow 0)$.

Felsenstein was concerned with the issue of *statistical consistency*; in this context, consistency refers to the conditions under which characters would recover the model tree $(AB|CD)$ as opposed to the alternatives $(AC|BD$ or $AD|BC)$. There are six character distributions relevant to this problem: two for each of the three alternate splits (Eq. 11.1), where the number of each characters supporting a split (n_{ABCD}) are:

$$\begin{aligned} AB|CD &: n_{1100} + n_{0011} \\ AC|BD &: n_{1010} + n_{0101} \\ AD|BC &: n_{1001} + n_{0110} \end{aligned} \tag{11.1}$$

Each of these conditions has an associated probability (starting from the root) based on p and q (Eq. 11.2):

$$\begin{aligned} pr_{1100} &= pq\left[(1-q)^2(1-p) + q^2p\right] \\ pr_{0011} &= (1-q)(1-p)\left[q(1-q)(1-p) + (1-q)pq\right] \\ pr_{1010} &= p(1-q)\left[q^2(1-p) + (1-q)^2p\right] \\ pr_{0101} &= (1-p)q\left[q(1-q)p + (1-q)q(1-p)\right] \\ pr_{1001} &= p(1-q)\left[q(1-q)p + (1-q)q(1-p)\right] \\ pr_{0110} &= (1-p)q\left[q^2(1-p) + (1-q)^2p\right] \end{aligned} \tag{11.2}$$

In order for the parsimonious result to return the model tree, the probability of those characters supporting the tree must be greater than that for the two alternatives (Eq. 11.3).

$$pr_{1100} + pr_{0011} \geq pr_{1010} + pr_{0101},\ pr_{1001} + pr_{0110} \tag{11.3}$$

If $q \leq \frac{1}{2}$ (which we assume), then $pr_{1010} + pr_{0101} \geq pr_{1001} + pr_{0110}$. Hence, the condition we require is that $pr_{1100} + pr_{0011} \geq pr_{1010} + pr_{0101}$. This will be achieved when the probability of two parallel changes in p exceeds that of a single change in q (Eq. 11.4).

$$p^2 \leq q(1-q) \tag{11.4}$$

The key relationship is between p and q. As long as p grows with respect to q, parsimony will be increasingly unlikely to return the model tree (Fig. 11.2)[1].

Criticisms and qualifications of this result are argued in discussions of the relative merits of optimality criteria and are discussed in Chapter 13.

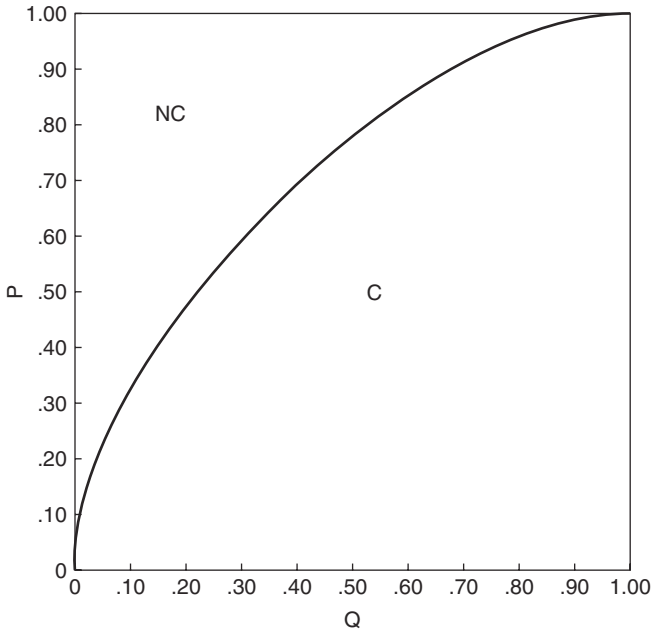


Figure 11.2: The "Felsenstein Zone" (NC) of statistical inconsistency of parsimony (Felsenstein, 1978).

[1]This effect is removed if p/q is constant and p and q become adequately small (Felsenstein, 1973) or the number of states increases sufficiently (Steel and Penny, 2000).

11.2 Maximum Likelihood and Trees

From the discussion of Section 6.1.7, the likelihood of a hypothesis (in this case a tree T) given data D, is proportional to the probability of the data given the tree (and some model; Eq. 11.5, Edwards, 1972).

$$l(T|D) \propto pr(D|T) \tag{11.5}$$

A systematic ML method selects T such that $pr(D|T)$ is maximized. This statement includes the requirement of knowledge of a broad variety of quantities needed to determine the likelihood. These include transformation models, edge distribution (branch lengths), and other parameters bundled together under the term "nuisance parameters."

11.2.1 Nuisance Parameters

Nuisance parameters are all those aspects required to calculate $pr(D|T)$ other than the data and tree topology. The three most important and commonly specified nuisance parameters are 1) transformation model (probabilities of change between character states), 2) edge parameters (time and rate of change along branches), and 3) distribution of rates of change among characters. These parameters can be denoted collectively by θ, and are estimated from observed data (as with edge parameters), or chosen to maximize the likelihood of a tree or trees. An important assumption for the analysis of character data is that they are *independent and identically distributed* (i.i.d.). This allows the joint likelihood of several characters to be calculated as the product of their individual values. Certainly, for many character types, this is not reasonable (*e.g.* stem and loop sequence characters in rRNA). However, distributional models can account for this to a large extent (although dynamic character types would be an exception).

If we have knowledge of the distribution of the nuisance parameters $\Phi(\theta|T)$, we can integrate out θ (within parameter space Θ) to determine $p(D|T)$ (Eq. 11.6).

$$p(D|T) = \int_{\theta \in \Theta} p(D|T,\theta) d\Phi(\theta|T) \tag{11.6}$$

That T which maximizes $p(D|T)$ in this way is referred to as the *maximum integrated likelihood* (MIL) (Steel and Penny, 2000). The MIL is also the MAP Bayesian estimate (Chapter 12) if the distribution of tree priors is uniform (flat).

When discussing stochastic model-based systematic methods, it can be useful to determine the probability that a given method, M, will return the "true" tree given a tree T and set of model parameters θ, $\rho(M,T,\theta)$. If we have $\Phi(\theta|T)$ and a prior distribution of trees, $p(T)$, the nuisance parameters and tree can be integrated out, identifying M with the highest expectation of success (Eq. 11.7).

$$\rho(M) = \sum_{T} p(T) \int_{\theta \in \Theta} \rho(M,T,\theta) d\Phi(\theta|T) \tag{11.7}$$

Székely and Steel (1999) showed that $\rho(M)$ is maximized for the method that returns T with maximum $p(T)pr(D|T)$. This is the Bayesian *maximum a posteriori* or MAP tree. As mentioned above, this is identical to the MIL tree when all prior probabilities of trees are equal. The use of non-uniform tree priors (such as empirical or Yule) breaks this identity.

11.3 Types of Likelihood

As mentioned above, θ can have many complex components, and we are unlikely to have much knowledge of their distribution. One approach to circumvent this problem is to choose θ such that $p(D|T,\theta)$ is maximized. This is referred to as *maximum relative likelihood* (MRL). In general, this is the methodology used in empirical analyses. Problems may arise when $p(D|T,\theta) > p(D|T',\theta')$ for a low probability θ (if we were to have $\Phi(\theta|T)$) while for a set of high probability θ, $p(D|T',\theta') > p(D|T,\theta)$. Steel and Penny (2000) cite such an example in a four-taxon case where parsimony outperforms MRL. MRL operates in absence of $p(T)$ and $\Phi(\theta|T)$, allowing likelihood analysis of systematic data. There are, however, further distinctions among MRL methods.

11.3.1 Flavors of Maximum Relative Likelihood


John Hartigan

There are three variants in the manner in which non-leaf character states are determined. The most usual method is to sum over all possible vertex state assignments weighted by their probabilities. In the nomenclature of Barry and Hartigan (1987), this is referred to as *maximum average likelihood* (MAL). An alternative would be to assign specific vertex states (as well as other parameters) such that the overall likelihood of the tree is maximized. Barry and Hartigan (1987) suggested this method, naming it *most parsimonious likelihood* (MPL, sometimes referred to as *ancestral maximum likelihood*). This would appear to be convergent with parsimony, but the edge probabilities are the same over all characters hence MPL will not (in general) choose the same tree as parsimony.

A third variant was proposed by Farris (1973a) and termed *evolutionary path likelihood* (EPL). In this form, the entire sequence of intermediate character states between vertices are specified such that the overall tree likelihood is maximized. Interestingly, the tree which maximizes this form of likelihood is precisely the most parsimonious tree. This result holds for a broad and robust set of assumptions (there is no requirement of low or homogeneous rates of character change for example). This would conflict with Felsenstein's assertion of ML methods being consistent and Farris' result that MP is an ML method. This seeming paradox is resolved when it is realized that the forms of likelihood discussed by Farris and Felsenstein (and the separate analogous MP = ML results of Goldman, 1990 and Tuffley and Steel, 1997) differ (Fig. 11.3). For the remainder of this discussion, when we talk of ML methods, we will be referring to MAL.

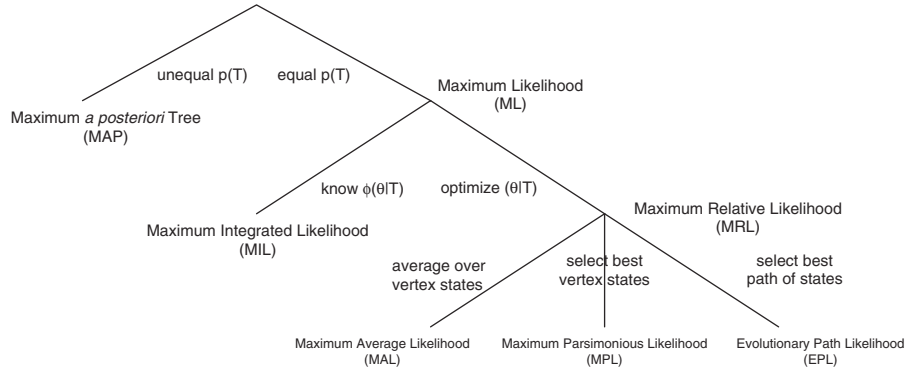


Figure 11.3: A classification of likelihood methods employed in systematics.

11.4 Static-Homology Characters

11.4.1 Models

Character Transformation

We can create a general model for a character of n states, with instantaneous transition (rate) parameters between states i and j, R_{ij}, and a vector of state frequencies Π (Eq. 11.8; Yang, 1994a).

$$R = \begin{bmatrix} R_{00} & \dots & R_{0n} \\ \cdot & \dots & \cdot \\ \cdot & \dots & \cdot \\ \cdot & \dots & \cdot \\ R_{n0} & \dots & R_{nn} \end{bmatrix} \Pi = \begin{bmatrix} \pi_0 \\ \cdot \\ \cdot \\ \cdot \\ \pi_n \end{bmatrix} \tag{11.8}$$

In general, we require several symmetry conditions of R (Eq. 11.9).

$$\begin{aligned} \forall i \; R_{ii} &= 0 \\ \forall i, j \; R_{ij} &= R_{ji} \\ \sum_{i=1}^{n} \sum_{j=1}^{n} \pi_i \cdot \pi_j \cdot R_{ij} &= 1 \end{aligned} \tag{11.9}$$

The combination of these two matrices yields the Q, or rate matrix, of Tavaré (1986) (Eq. 11.10).

$$Q_{ij} = \begin{cases} R_{i,j} \cdot \pi_j & i \neq j \\ -\sum_{m=1}^{n} R_{i,m} \cdot \pi_m & i = j \end{cases} \tag{11.10}$$

The probability of change (P) between states i and j in time t can be calculated from elementary linear algebra (Eq. 11.11; Sect. 6.2):

$$P_{i,j}(t) = \sum_{m=1}^{n} e^{\lambda_m t} \cdot U_{m,i} \cdot U_{j,m}^{-1} \tag{11.11}$$

with λ_m, the eigenvalues of Q, U the associated matrix of eigenvectors and U^{-1}, its inverse (Strang, 2006). The time-reversible constraint of the matrix allows efficient computation of tree likelihoods.

This formulation is the most general (if symmetrical) description of a Markov process for n character states. This model has, at most, $n-1$ independent frequency parameters (Π; one for each state, but the total must sum to 1) and $\binom{n}{2} - 1$ independent rate parameters (R) due to the constraints above (Eq. 11.9).

Special Cases

All character transformation models in use today, from the simple binary model of Felsenstein (1973), through the four state homogeneous Jukes and Cantor (1969) to General-Time-Reversible models for four (Lanave et al., 1984; Tavaré, 1986) and five states (McGuire et al., 2001; Wheeler, 2006), are simplifications of the most general process through symmetry requirements (*e.g.* transversions equal). All of the named models other than GTR (*e.g.* JC69) are special cases where analytical solutions are known (as opposed to computationally determining eigenvalues and applying Eq. 11.11). The hierarchy of simplifications for four states is illustrated in Swofford et al. (1996) (Fig. 11.4).

11.4.2 Rate Variation

In addition to models of character transformation, there are also distributional models of character change rates. These are most frequently used in the analysis of molecular sequence data where aligned nucleotide characters are analyzed as

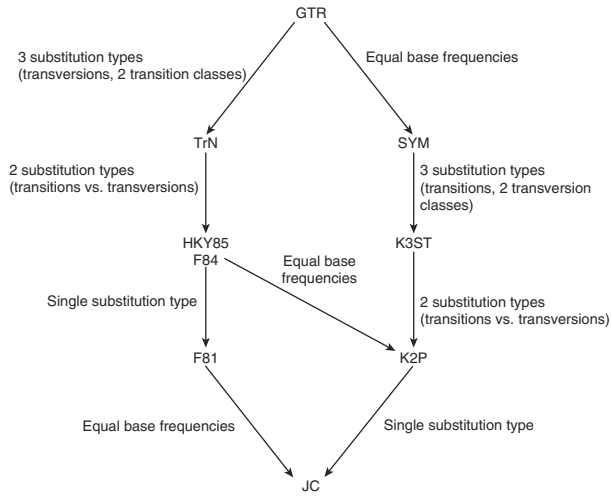


Figure 11.4: Swofford et al. (1996) relationships among DNA substitution likelihood models from the least parameterized JC69 to the most, GTR.

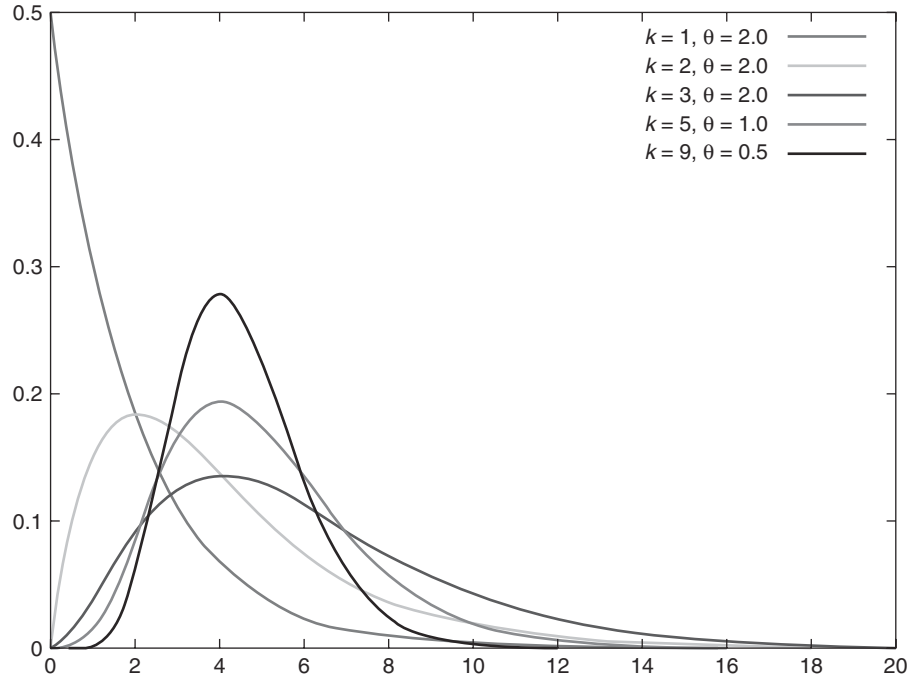


Figure 11.5: The gamma distribution with shape parameter $\alpha = k$, $\beta = \theta$.

a block. Positions vary in their observed levels of variation (hence, evolutionary rates), and this is accommodated by adding variation to the global rates of change used to calculate tree likelihoods.

The two most common are the fraction of invariant sites (Hasegawa et al., 1985) and discrete-gamma distribution (Yang, 1994a). The notion behind the use of an invariant sites parameter (usually referred to by I) is that one frequently observes many invariant positions with sequence data and accounting for this class of positions with a global rate is undesirable. Hence, a parameter is added to account for the fraction of sites available for substitution.

The gamma distribution (used in its computable discrete form), adds additional classes of positional rates based on a shape parameter α. The distribution (Eq. 11.12, Fig. 11.5) has a mean of α/β and variance of α/β^2, but we usually set $\beta = \alpha$ for a mean of 1.

$$g(x; \alpha, \beta) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \tag{11.12}$$

The user specifies a number of rate classes (often in concert with invariant sites) and estimates α such that the tree likelihood is maximized. It is important to note that all rate classes are applied to each position, as opposed to a single class to a given position. For n taxa, m characters (*e.g.* aligned nucleotide sites), s states, and r rate classes, the overall memory consumption will be $O(nmsr)$.

11.4.3 Calculating $p(D|T,\theta)$

For a single character (x) on a tree, the likelihood of internal vertex i (L_i) with descendant vertices j and k would be the sum of the probability between x_i and each state in each descendant (given the edge parameter t; Fig. 11.6) multiplied by its respective likelihood and summed over all states. The character likelihoods are multiplied over the entire data set to determine the tree likelihood (Eq. 11.13).

$$L_i(x) = \sum_{i}^{states} \left[\left(\sum_{x_j} p_{x_i,x_j}(t_j) L_j(x_j) \right) \times \left(\sum_{x_k} p_{x_i,x_k}(t_k) L_k(x_k) \right) \right] \quad (11.13)$$

When edge weights are not known (as in nearly all real data situations), they must be estimated. This can be done in several ways, but all rely on calculation of the marginal likelihood (holding all other parameters constant) of a given edge assuming a variety of weights (t parameter) and choosing the optimal value (Fig. 11.7). Often Brent's Method (Brent, 1973) or Newton–Raphson (Ypma, 1995) is used to estimate the edge parameters.

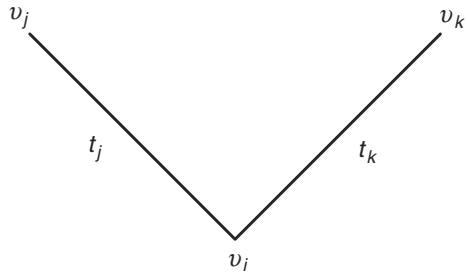


Figure 11.6: Labeled subtree for likelihood calculations.

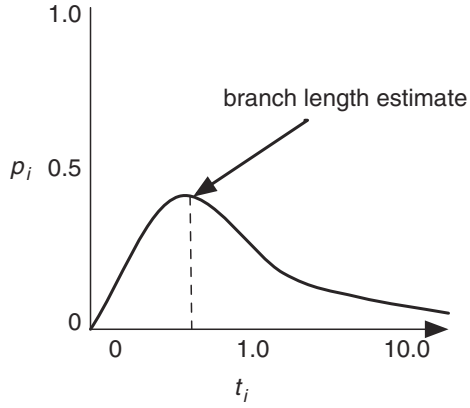


Figure 11.7: Estimate of edge weight parameter t by maximizing the probability of transformation along the edge, p_i.

The determination of the MAL of a tree is a heuristic procedure due to the large number of parameter estimations involved. As with parsimony optimization (Chapter 10), the tree is traversed setting median states recursively. This recursion is initialized with the likelihood of leaf states at 1 (no need to sum to one for likelihood) and all other leaf states 0. The likelihood is calculated via a post-order tree traversal from the tips to the root multiplied by the prior probabilities of the states themselves (Eq. 11.14).

$$L_T(x) = \prod_{i=1}^{states} \pi_i \prod_{\forall u,v \in E} L_{u,v} \tag{11.14}$$

Given that these values can be quite small, it is often convenient to speak of log or $-\log$ likelihood values[2]. The following example assumes that the edge parameters are known. If this is not so (which is usually the case), such a single post-order traversal will not be sufficient to determine the tree likelihood. An iterative edge refinement procedure will be required to optimize the edge parameters (Felsenstein, 1981).

An Example

Consider a single nucleotide character analyzed under the JC69 model (Fig. 11.9). If we fix all the edge probabilities, $\mu t = 0.1$, we can calculate the likelihood of the topology given the analytical probabilities in Equation 11.15.

$$P_{ij}(t) = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-\mu t} & i = j \\ \\ \frac{1}{4} - \frac{1}{4}e^{-\mu t} & i \neq j \end{cases} \tag{11.15}$$

Hence, the edge probabilities are given in Equation 11.16.

$$P_{ij}(t) = \begin{cases} 0.929 & i = j \\ \\ 0.0238 & i \neq j \end{cases} \tag{11.16}$$

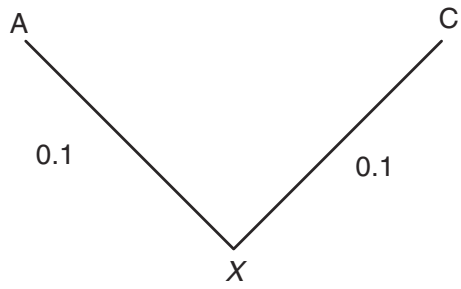
A subtree example with leaf states (A and C) and edge parameters 0.1 is shown in Figure 11.8.

The overall likelihood for the tree in Figure 11.9 is 1.76×10^{-6} or, in familiar $-\log$ (base e) units, 13.25.

11.4.4 Links Between Likelihood and Parsimony

Typical likelihood analyses employ several homogeneity conditions. Usually the same edge parameter is applied to all characters (although it may vary over

[2]The finite precision of computers can cause problems for likelihood calculations (floating point error) due to the large number of operations required when evaluating trees. Alternate implementations of the same algorithm may well generate likelihoods that differ non-trivially. Extreme care must be taken to avoid this problem.



$$L(x = \mathrm{A}) = [0.929 \cdot 1.0] \times [0.0238 \cdot 1.0] = 0.0221$$
$$L(x = \mathrm{C}) = [0.0238 \cdot 1.0] \times [0.929 \cdot 1.0] = 0.0221$$
$$L(x = \mathrm{G}) = [0.0238 \cdot 1.0] \times [0.0238 \cdot 1.0] = 0.000566$$
$$L(x = \mathrm{T}) = [0.0238 \cdot 1.0] \times [0.0238 \cdot 1.0] = 0.000566$$
$$\text{Total } L(x) = 0.0453$$

Figure 11.8: Labeled subtree with likelihood calculations.

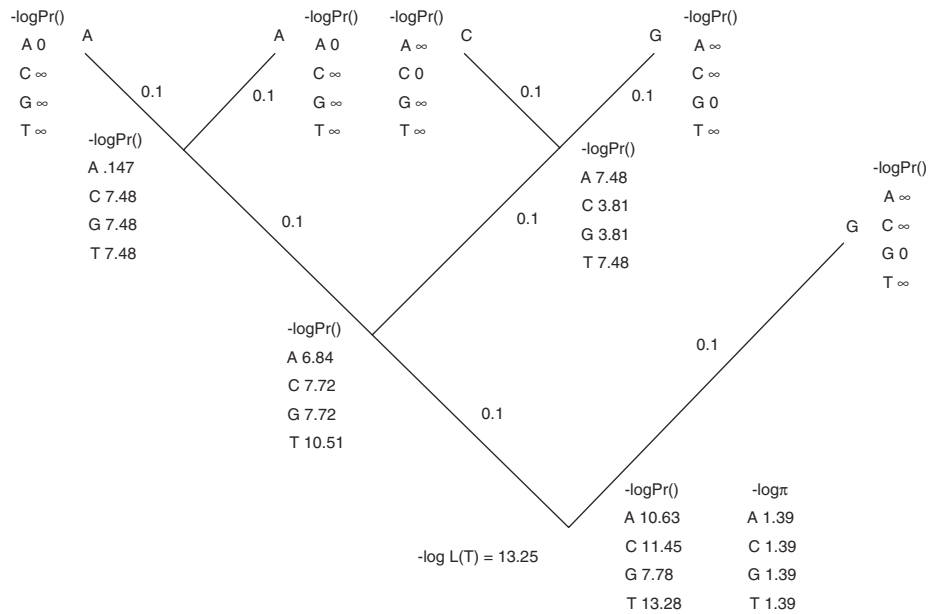


Figure 11.9: An example likelihood calculation under JC69 model with all edge parameters set to 0.1.


Nicholas Goldman


Chistopher Tuffley

edges), and the same model as well. Under these conditions, MAL will frequently lead to results at variance with parsimony. As mentioned earlier, Farris (1973a) employed a simple model to show that parsimony and EPL would choose the same optimal tree. Such connections are not limited to this scenario.

Goldman (1990) discussed a number of scenarios involving likelihood, parsimony, and compatibility. Goldman showed that when edge weights are constant over the tree, likelihood and parsimony will converge. More recently, Tuffley and Steel (1997) discussed the No-Common-Mechanism (NCM) model, where each character has a potentially unique rate that may vary among edges as well (hence the name). The rate for each character on each edge is optimized (either zero, or infinite) to maximize the likelihood. Under a Neyman (1971) type model with r states, the overall likelihood of the tree can be determined as a function of the number and distribution of parsimony changes on the tree (Eq. 11.17), with r_i, the number of states exhibited by character i, χ_i, the parsimonious vertex states assignments for character i, and $-l(\chi_i, T)$, the parsimony length of assignment χ for character i on tree T.

$$L_T(X) = \prod_{i=1}^{k_{characters}} r_i^{-l(\chi_i,T)-1} \tag{11.17}$$

For the tree and leaf states of Figure 11.9, the likelihood would be $3^{-(2+1)} = 0.037$. It is often said that this model leads to equivalent results between parsimony and likelihood, but this will only occur when the number of states of each character (r_i) is a constant over the data set. In this way, NCM can be viewed as a likelihood-based character weighting scheme in parsimony analyses.

11.4.5 A Note on Missing Data

Missing data are not, in principle, a problem for likelihood analyses. Leaf state vectors can be set to 1.0 for each of the observed (in the case of polymorphism) or implied (all states = 1.0) states in the case of entirely missing observations. Implementations, however, may differ in the treatment of these unknown observations. Currently, implementations treat missing data in this manner. Obviously, this can have an effect on analyses. This issue can become all the more pernicious when coupled with the practice of treating indels or "gap" characters as missing values (as opposed to a 5th state). Though clearly suboptimal (and unnecessary, as shown later), such a treatment of indels is common and problematic.

11.5 Dynamic-Homology Characters

As with parsimony, maximum likelihood can be applied to the analysis of dynamic-homology characters. With sequence (nucleotide and amino-acid) and higher order characters (*e.g.* gene rearrangement), two general approaches have been taken in the construction of stochastic models. The first uses a simple

extension of 4-state nucleotide or 20-state amino-acid models to include "gaps" as a 5th or 21st state (Wheeler, 2006). These models treat indels as atomic events, emphasize simplicity and make little attempt to model reality *per se*. The second approach makes an explicit attempt to model the process of sequence change including indel events (Thorne et al., 1991, 1992), resulting in more complex scenarios.

In general, models describing the process of gene rearrangement are not attempts to describe the mechanisms of genomic change as much as descriptive statements of the frequency and patterns of change (*e.g.* Larget et al., 2004).

11.5.1 Sequence Characters

In order to perform a dynamic homology analysis (Tree Alignment Problem; Chapter 10) of multiple leaf sequences related by a tree, several components are required. First, a model must be specified allowing both element substitution and insertion–deletion (indel). Second, a procedure needs to be identified to calculate the likelihood "distance" between any pair of sequences. And third, a method of creating sequence medians (vertex or HTU sequences) must be described.

Models

$n+1$ *State Models*—A simple expansion of sequence substitution models to include an extra state for "gaps" representing indels (such as the r-state model of Neyman, 1971) has been used by McGuire et al. (2001) in their Bayesian analysis of pre-aligned sequences and the ML Direct Optimization (ML–DO) of Wheeler (2006).

In the symmetrical ($r_{ij} = r_{ji}$, R of Tavaré, 1986) general 5-state case there are five state frequencies to be specified (A, C, G, T, -), although they must sum to 1, and 10 transition rates among the states (Fig. 11.10). As with the GTR model of sequence substitution above, there are a broad variety of special case models that can be constructed by enforcing various additional symmetry conditions (such as JC69+Gaps, Eq. 11.18; Wheeler, 2006).

$$P_{ij}(t) = \begin{cases} \frac{1}{5} + \frac{4}{5}e^{-\mu t} & i = j \\ \\ \frac{1}{5} - \frac{1}{5}e^{-\mu t} & i \neq j \end{cases} \tag{11.18}$$

Considering the example alignment of 11.19 under the model in Equation 11.18 with an edge weight (branch length) of 0.1 (μt), $p(I, II) = (0.01903)^3(0.9239)^2 = 5.882 \times 10^{-6}$.

$$\begin{array}{rl} \text{Sequence I} & \text{AC-GT} \\ \text{Sequence II} & \text{AGC-T} \end{array} \tag{11.19}$$

	A	C	G	T	-
A	$-(\pi_C\alpha+\pi_G\beta+\pi_T\gamma+\pi\_\delta)$	$\pi_C\alpha$	$\pi_G\beta$	$\pi_T\gamma$	π_δ
C	$\pi_A\alpha$	$-(\pi_A\alpha+\pi_G\varepsilon+\pi_T\zeta+\pi\_\eta)$	$\pi_G\varepsilon$	$\pi_T\zeta$	π_η
G	$\pi_A\beta$	$\pi_C\varepsilon$	$-(\pi_A\beta+\pi_C\varepsilon+\pi_T\theta+\pi\_\kappa)$	$\pi_T\theta$	π_κ
T	$\pi_A\gamma$	$\pi_C\zeta$	$\pi_G\theta$	$-(\pi_A\gamma+\pi_C\zeta+\pi_G\theta+\pi\_\nu)$	π_ν
-	$\pi_A\delta$	$\pi_C\eta$	$\pi_G\kappa$	$\pi_T\nu$	$-(\pi_A\delta+\pi_C\eta+\pi_G\kappa+\pi_T\nu)$

Figure 11.10: A general, symmetrical, 5-state model (states A, C, G, T, '-').

These models have the virtue of simplicity and ease of calculation, hence can be applied to real data sets with multiple loci and empirically interesting (>100) numbers of taxa (Whiting et al., 2006).

Birth–Death Model—The Thorne et al. (1991) and Thorne et al. (1992) models (TKF91 and TKF92), treat the insertion–deletion process in an alternate fashion. There are two components to the calculation of the probability of transforming one sequence into another: the probability of an alignment (α as in 11.19) given a set of insertions, deletions, and matches and model $[p(\alpha|\alpha',\theta)]$; and the probability of a specific pattern of indels and matches given a model $[p(\alpha'|\theta)]$. The method couples a birth–death process (parameters λ—insertion or birth rate; μ—deletion or death rate) with standard four-nucleotide substitution models.

Both TKF91 and TKF92 model the indel process in the same way, transforming one sequence into another (the model is symmetrical). There are three sorts of events. The first is an insertion (not leading) in the first sequence to yield the second (p). The second transformation type is a deletion (p'), and the third, a leading insertion, takes place before the left-most residue (p''). The probabilities of these structural events are as in Equation 11.20, with λ birth rate (insertion), μ death rate (deletion), $n > 0$ indel size, and time t.

$$\begin{aligned} p_n(t) &= e^{-\mu t}\left[1-\lambda\beta(t)\right]\left[\lambda\beta(t)\right]^{n-1} \\ p'_n(t) &= \left[1-e^{-\mu t}-\mu\beta(t)\right]\left[1-\lambda\beta(t)\right]\left[\lambda\beta(t)\right]^{n-1} \\ p''_n(t) &= \left[1-\lambda\beta(t)\right]\left[\lambda\beta(t)\right]^{n-1} \end{aligned} \tag{11.20}$$

with

$$\beta(t) = \frac{1-e^{(\lambda-\mu)t}}{\mu-\lambda e^{(\lambda-\mu)t}}$$

The substitution process follows standard models with state frequencies determining the probability of inserting a specific sequence.

11.5.2 Calculating ML Pairwise Alignment

Both the above models can be optimized for two sequences by versions of the familiar dynamic programming procedure used for pairwise sequence alignment (Sect. 8.4). Here, we discuss the algorithm for the $n + 1$ state model. The recursions are more complex for TKF92, but they follow the same basic outline (see Thorne et al., 1992 for specifics).

Dynamic Programming

In order to calculate the probability of transforming one sequence into another (or a pairwise alignment; as with parsimony the cost is identical for two sequences), three elements are required: the sequences, the transformational model, and a time parameter to mark the differentiation between the sequences $[p(I, II|\theta, \tau)]$. Dynamic programming will optimize the likelihood for a given t, but as with edge weight/branch length optimization, the procedure must be repeated, varying or estimating t until the likelihood is optimized (Eq. 11.21).

$$p(I, II|\theta) = \max_t p(I, II|\theta) \tag{11.21}$$

Since t is chosen to maximize the pairwise probability, the method will yield an MRL.

It is often convenient to work with the negative logarithm of likelihood and probability values as opposed to their absolute values and, in this case, it allows an elegant modification of the Needleman and Wunsch (1970) algorithm (Alg. 8.1). Based on model and time, the conditional probability of an indel or element match can be calculated *a priori*. In the scenario above (JC69+Gaps with $t = 0.1$), the probability of an indel is $\frac{1}{5} - \frac{1}{5}e^{-0.1} = 0.01903$, an element mismatch (substitution) is the same $\frac{1}{5} - \frac{1}{5}e^{-0.1} = 0.01903$, while an element match $\frac{1}{5} + \frac{4}{5}e^{-0.1} = 0.9239$. Using the logarithms of these values, the multiplicative probabilities of a scenario can be optimized as additive sums $[\log(p(x_i) \cdot p(x_j)) \rightarrow \log p(x_i) + \log p(x_j)]$ by treating them as match, mismatch, and indel costs.

Although the log transform probabilities can be used as edit costs (*i.e.* $cost[i][j] = \log p(I_i, II_j|\theta, t)$), the core recursion requires a modification. The probability of inter-transforming (or aligning) two sequences is the sum of the probabilities of all potential transformation (or alignment) scenarios between the two. As we know (Eq. 8.6; Slowinski, 1998), there are a large number of these to calculate. The Needleman–Wunsch algorithm can accomplish this when the central alignment recursion is changed to a sum as opposed to the minimum of three paths. This sum is taken among the probabilities (*not* log probabilities) of the three options (element insertion, deletion, and match) at each cell (Eq. 11.22) ($cost[i][j]$ is the $-\log$ transformed likelihood).

$$\begin{aligned} cost[i][j] \;=\; & \log(e^{-(cost[i-1][j-1]+\sigma_{i,j})} \\ & + e^{-(cost[i-1][j]+\sigma_{indel})} \\ & + e^{-(cost[i][j-1]+\sigma_{indel})}) \end{aligned} \tag{11.22}$$

For sequences of lengths n and m, $p(I, II|\theta, t) = e^{-cost[n][m]}$. The traceback diagonal marks the maximum likelihood path as before. The complete matrix (in $\log_e$ units) is shown in Figure 11.11 resulting in a $p(I, II|\theta = JC69{+}Gaps, t = 0.1) = 0.0007849$. If one were to calculate the probability directly from the four aligned positions the value would be $0.01903^2 \cdot 0.8535^2 = 0.0002638$, considerably lower than that yielded by the algorithm. This is because the specific

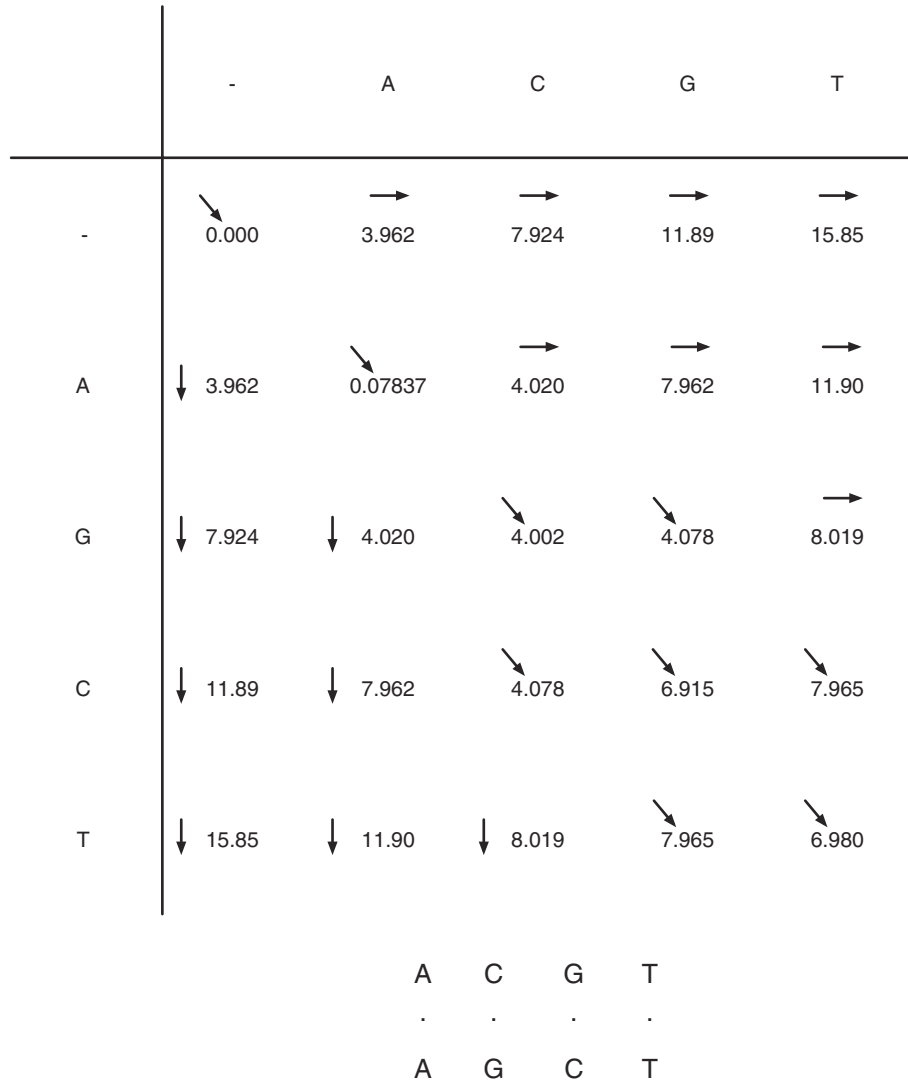


Figure 11.11: Likelihood alignment of two sequences (ACGT and AGCT) under the JC69+Gaps (5-state Neyman) model with a time parameter (μt) of 0.1 ($\log_e$ units).

alignment produced is only one of many alignment scenarios that contribute to the total probability of transformation between the sequences. This particular alignment has the highest probability of all possible alignments, hence is termed the *dominant* likelihood alignment (in the terminology of Thorne et al., 1991). We can search directly for this by choosing the maximum probability choice (insertion, deletion, or element match) in Eq. 11.22 as opposed to the sum (Fig. 11.12). The probability produced in this way jibes precisely with that

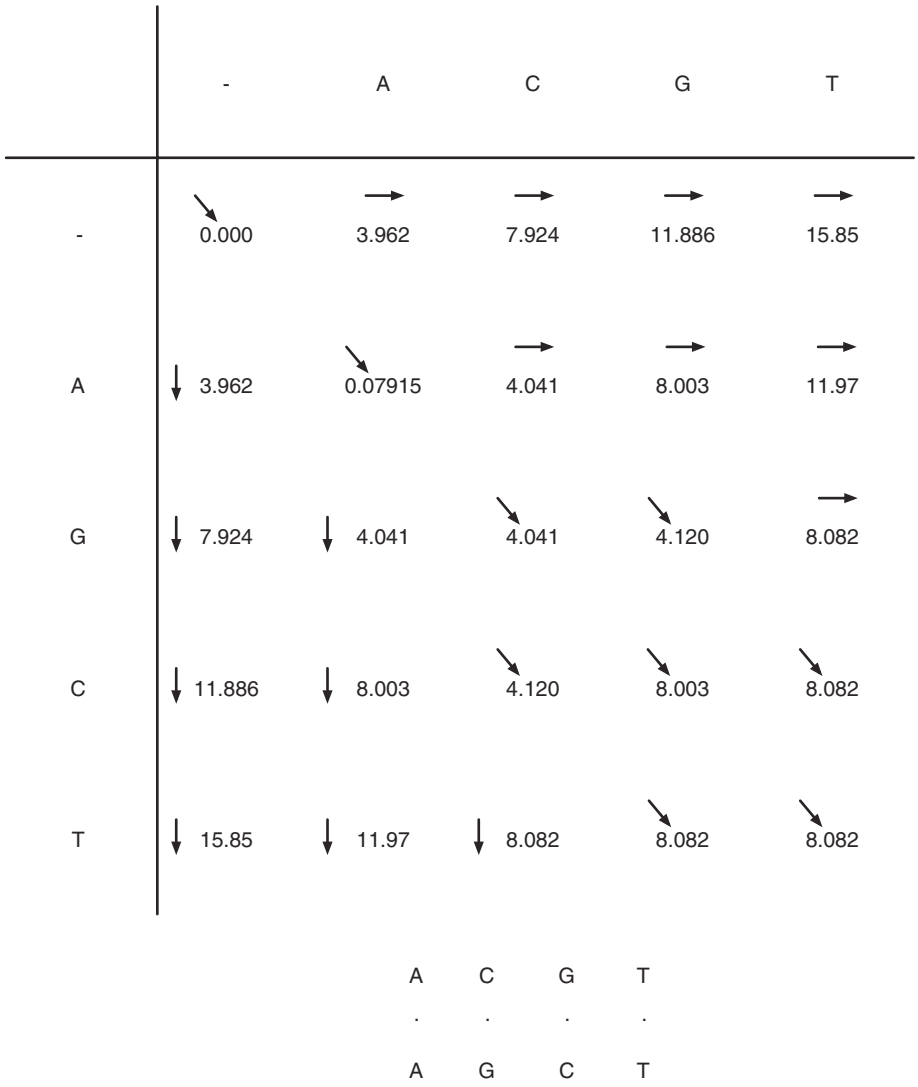


Figure 11.12: Dominant likelihood alignment of two sequences (ACGT and AGCT) under the JC69+Gaps (5-state Neyman) model with a time parameter (μt) of 0.1 ($\log_e$ units).

expected ($e^{-8.082} = 0.0003091$). In this case, both procedures yielded the same alignment, but this need not be the case in general.

The distinction between *dominant* and *total* likelihood is an important one. A single alignment may be "best" in a likelihood context, but may contain a very small fraction of the total likelihood (in this case 33%). In the context of likelihood forms discussed above, the dominant likelihood is akin to an MPL object, and the total likelihood score MAL. When sequence change is analyzed on a tree, these distinctions have downstream ramifications in the identification of ML trees, and character change maps on those trees.

11.5.3 ML Multiple Alignment

As with parsimony, there are relatively direct extensions of pairwise alignment to multiple sequence alignment (MSA). The approach of Wheeler (2006) was to create an implied alignment (Wheeler, 2003a) using the maximum likelihood form of Direct Optimization (Wheeler, 1996). In this ML–TAP approach, medians (and tree topologies) are chosen to optimize likelihood under a variety of models from a 5-state Neyman scenario to an enhanced GTR+Gaps model. The relative performance of parsimony and ML implied alignments was tested by Whiting et al. (2006), showing (comfortingly) that ML MSA were superior for ML (by 10% log likelihood units) while those based on parsimony were superior for parsimony analysis (by 30%; manual alignments were distant finishers; Table 11.1).

MSA methods based on the TKY92 model (Thorne et al., 1992), such as Fleissner et al. (2005) and Redelings and Suchard (2005), make use of Bayesian Hidden Markov Models and are discussed briefly above and in Chapter 12 in more detail.

11.5.4 Maximum Likelihood Tree Alignment Problem

Although it is as yet unstudied, given the NP–hard nature of the parsimony version of the TAP, the ML variant is likely to be extremely challenging if not NP–hard itself. As with parsimony heuristics to the TAP, we can generate several heuristic ML–TAP procedures. Unfortunately, almost nothing is known about the quality of these solutions (boundedness).

	ClustalX	Manual	DO–MP	DO–ML
Mixed Model Likelihood	61,489.630	55,329.945	51,611.928	50,496.073
Single Model Likelihood	61,548.268	55,858.397	51,554.225	51,014.655
Parsimony Tree Length	15,154	20,341	11,483	11,702

Table 11.1: Performance of ClustalX (Higgins and Sharp, 1988), Manual, DO–MP, and DO–ML multiple sequence alignment (Whiting et al., 2006). DO implied alignment runs were created using Wheeler et al. (2005) and ML scores by Huelsenbeck and Ronquist (2003).

Medians and Edges

As with parsimony heuristics to the TAP, the identification of median sequences is crucial to the quality of the solution. ML–TAP has the added factor of edge or branch time. As mentioned above for the two sequence case, the probability of the alignment is dependent on the time parameter that is identified numerically through repeated (or estimated) likelihood optimizations.

Sections 10.9.2, 10.9.3, and 10.9.4 identified sequence medians in ways that are directly applicable to ML. The algorithm for determining sequence medians using Direct Optimization (Alg. 10.7; DO, Wheeler, 1996) can be applied largely without modification. Two issues merit attention. The first is the use of dominant or total likelihood for tree likelihood values and medians. Total likelihood will reflect more of the probability of alternate medians, but unless these medians are of optimal (in this case highest probability) cost, they will not be reflected in the median sequences. Dominant likelihood calculations maintain a more consistent approach in that the tree likelihoods are directly traceable to these specific sequences. When the total likelihood is used, this connection can be lost (Wheeler, 2006). The second issue centers around the median determination and time parameter. The time parameter interacts with the median identification process, not only to determine the probability of an ancestor–descendant transformation, but the ancestral sequences (= medians) themselves. When edge times are estimated to optimize likelihood scores, the medians themselves are likely to change, creating additional time complexity in the process. This is especially prominent when using iterative improvement methods (Sankoff and Cedergren, 1983; Wheeler, 2003b, 2006). With iterative improvement, there are three edges incident on a vertex which require simultaneous optimization in addition to the 3-dimensional median calculation.

Lifted, Fixed-States, and Search-Based (Sect. 10.9.3) procedures deal with a fixed pool of medians, hence that component of time complexity is reduced. Edge iteration is still an issue in two ways. First, the pairwise probability of transformation between states is time dependent. This can be either held constant over all state pairs, or be optimized (in a fashion akin to Tuffley and Steel, 1997) uniquely for each sequence pair. Secondly, edge times can be applied while a tree is optimized (using a single time) or using optimized times from the pairwise sequence comparisons.

The issue of dominant and total likelihood also enters in this class of heuristics through the summing (as in average likelihood) over all potential sequence medians, or the identification of the most likely medians (MPL) and determining tree likelihood on that basis.

Wheeler (2006) discussed the above ML–TAP heuristics in the context of a 5-state model, although they could be applied to other models. Fleissner et al. (2005) developed a heuristic ML–TAP procedure specifically for the TKF92 model. In their approach, simulated annealing (Sect. 14.7) is used in two ways alternately. The first is to optimize the analytical parameters (substitution model parameters, indel birth and death values, fragment length), and the second to optimize the alignment patterns of indels (h, α' of Thorne et al., 1991) and tree

topology. The method initializes with a Neighbor-Joining (Saitou and Nei, 1987) tree and performs NNI (Sect. 14.3.2) to break out of local topological minima. The parameter and topology/indel pattern optimizations proceed alternately until improvements are no longer found. Due to the complexity of the TKF92 model and the simulated annealing approach, the method does not scale well and can only be used on a handful (<20) of sequences of moderate (< 500bp) length.

11.5.5 Genomic Rearrangement

As with all stochastic procedures, the root of likelihood-based reconstruction of genomic rearrangement data is the model. Currently, models are descriptive, that is, distributions of gene rearrangements are chosen and fit to empirical patterns, not based on any first principles analysis of the biological mechanism of inversion or transposition (an exception exists in the Birth–Death model of gene family evolution of Zhang and Gu, 2004).

$$p(k, \lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{11.23}$$

The basic descriptive model was set out by Nadeau and Taylor (1984), grounded in the empirical observation of the distribution of chromosomal rearrangements between humans and mice (Fig. 11.13). They posited a Poisson distribution (Eq. 11.23) of rearrangement events (k) on the genome and along a tree edge at average rate λ. This was expanded by Wang and Warnow (2001, 2005) to create corrected distances for use in distance-based phylogenetic analysis (Chapter 9).

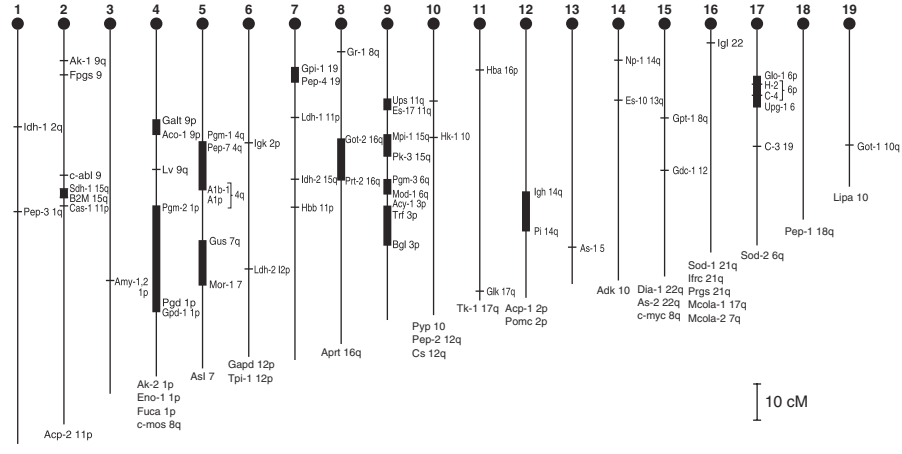


Figure 11.13: Mouse—Human rearrangements as illustrated by Nadeau and Taylor (1984).

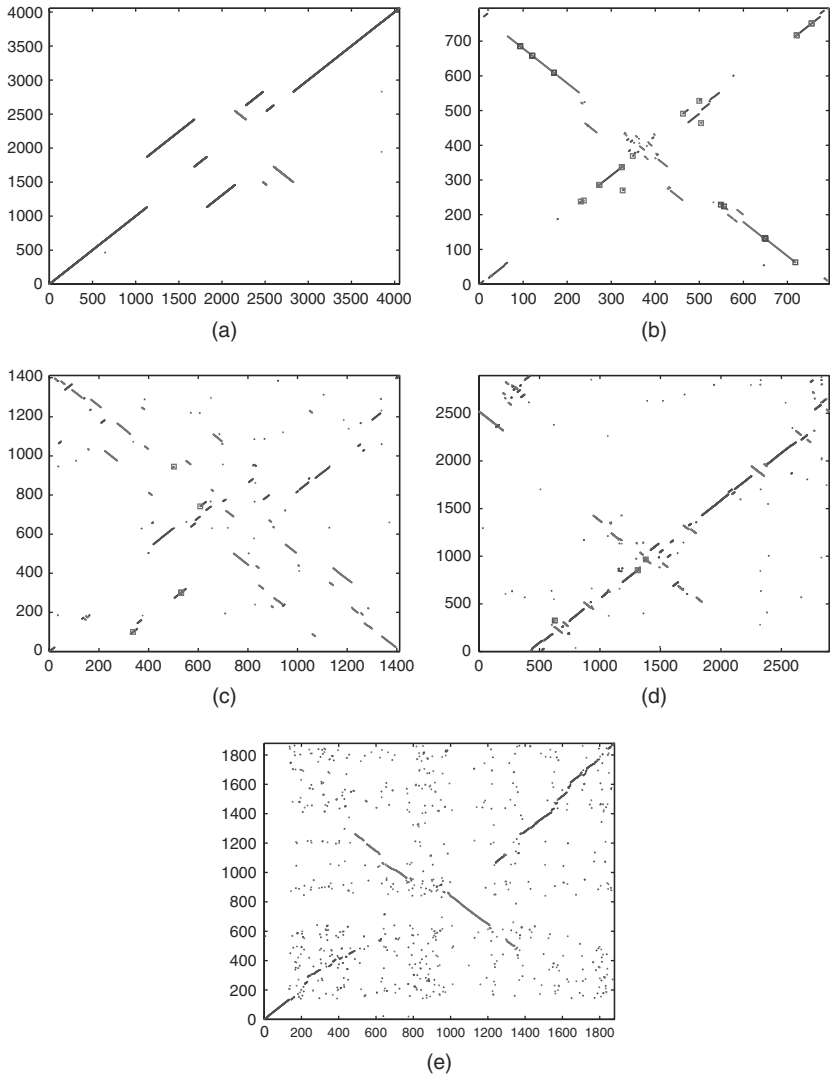


Figure 11.14: Genomic rearrangement locus dot-plot scenarios of Dalevi and Eriksen (2008): (a) = "Whirl," (b) = "X-model," (c) = "Fat X-model," (d) = "Zipper," and (e) = "Cloud." See Plate 11.14 for the color Figure.

Empirical Models

Dalevi and Eriksen (2008) presented a series of corrected distance estimates for five rearrangement scenarios named according to patterns on pairwise dot-plots (Fig. 11.14).

(a) "Whirl"—caused by an overrepresentation of uniformly distributed reversals across the genomes.

(b) "X-model"—due to a preponderance of reversals symmetrically distributed around the origins and terminations of replication.

(c) "Fat X-model"—explained by symmetrically distributed reversals with enhanced variation in their position with respect to the origins and terminations of replication.

(d) "Zipper"—thought to result from a large amount of short reversals (up to 5% of the genome) distributed uniformly over the genome.

(e) "Cloud"—as rearrangements accrue, the gene order becomes randomized loosing the previous patterns into a "cloud."

In general, these descriptions of rearrangement patterns are not used to reconstruct trees directly, but to estimate overall dissimilarity for distance analysis.

11.5.6 Phylogenetic Networks

As with parsimony (Sect. 10.14), horizontal gene transfer and hybridization can be explained by networks and in an analogous fashion (Strimmer and Moulton, 2000). Jin et al. (2006) proposed no biological model of horizontal gene transfer or hybridization, but two methods to calculate the likelihood of the network. In the same way that the parsimony score of a network is calculated by summing the minimum tree costs consistent with the network (Eq. 10.9) for each block of characters, likelihoods can be multiplied over the best likelihood tree for each character block. A second option would be to sum the likelihoods of all tree scenarios consistent with the network (Fig. 11.15; Eq. 11.24).

$$N = (V, E) \tag{11.24}$$

$$L_N^{all}(S|N,\theta) = \sum_{T \in N} \left(p(T) \cdot L(S|T,\theta)\right)$$

$$L_N^{best}(S|N,\theta) = \max_{T \in N} \left(p(T) \cdot L(S|T,\theta)\right)$$

It is unclear which, if either, procedure is appropriate. The first method assumes all blocks are independent. This may or may not be reasonable given that the recognition of blocks is dependent on their relative positions and behavior. The second model has the advantage of including alternate scenarios, weighted by their likelihoods, but allows for multiple histories for all blocks.

11.6 Hypothesis Testing

11.6.1 Likelihood Ratios

Often, it is desirable to know whether a difference between two likelihood values is "significant." As odd as such a concept may seem within the rationale of

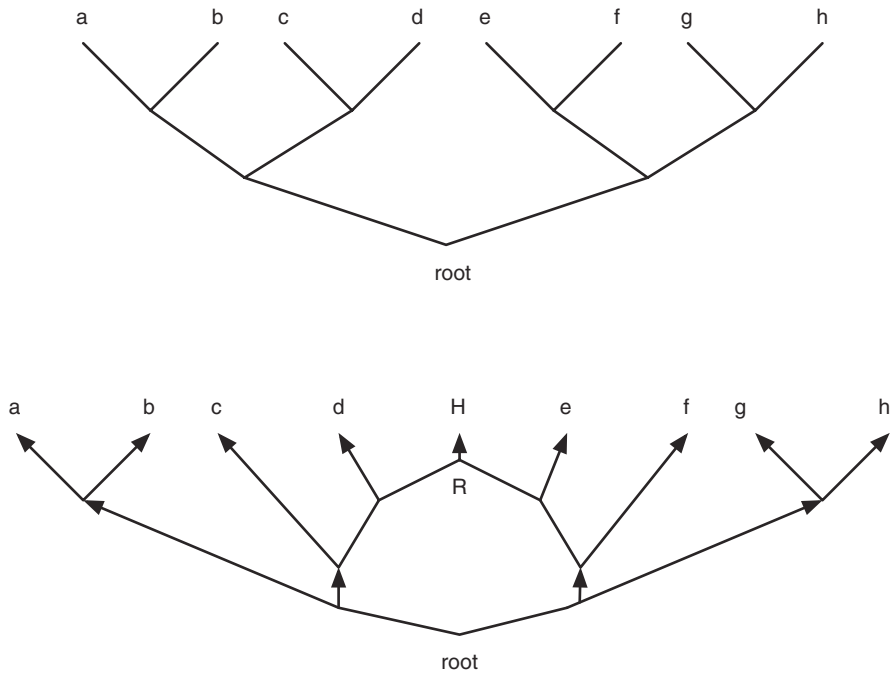


Figure 11.15: Phylogenetic tree (above) and network (below).

likelihood, given a few simple assumptions, such statements can be made (DeGroot and Schervish, 2006). The basic idea is that if the likelihood functions are well-behaved, twice the difference in the log of the ML value is distributed as χ^2 (Eq. 11.25). When the hypotheses to be compared are simple estimates of parameters (such as a branch length or comparison of two trees), this distribution will have one degree of freedom.

$$2\Delta l_{T,T'} = 2\log(l_T/l_{T'}) = 2(\log l_T - \log l_{T'}) \tag{11.25}$$

Likelihood ratio tests are used to determine whether edge weights (time parameter) are greater than 0 and should be collapsed, or whether one of two competing and nearly optimal hypotheses is superior. The confidence value is (given the single degree of freedom) 1.9207 log likelihood units, or a likelihood ratio of 6.826. If two tree likelihoods are differ by at least this value, their difference is statistically significant (Felsenstein, 2004).

Branch Collapsing

The likelihood ratio can also be used to test if an edge probability (ML branch length) is significantly greater than zero. The likelihood of a tree with an edge constrained to have $\mu t = 0$ can be determined and compared to the likelihood

of the tree optimized for the time parameter of that edge. As above (Eq. 11.25), the likelihood ratio can be tested via χ-squared with a single degree of freedom.

11.6.2 Parameters and Fit

As with all statistical fitting operations, increasing the number of parameters will increase the fit and decrease the error. In general, if the addition of a parameter results in a large increase in fit (here likelihood), we accept that parameter. The problem comes as more and more parameters are added and the increases in quality of solution (in terms of error) are less dramatic, leading to overparameterization and loss of predictivity (Fig. 11.16). An analysis of molecular sequence data using the JC69 model is based on zero parameters (everything is equal and nothing specified)[3]. The same data modeled using GTR would no doubt yield a better likelihood score using its eight parameters (five rate and three frequency). This might be further improved with invariant sites and discrete-gamma rate parameters. When should this stop? How can overparameterization be avoided?

There are two commonly used statistics to decide this. The first is the ratio of the likelihoods of solutions with different parameterization—the likelihood ratio test above. In the case of testing models, Equation 11.25 is distributed as $\chi^2_{p'-p}$ where p and p' are the number of parameters in the models to be compared.

Large sequence data sets nearly invariably choose the most complex models (GTR+I+Γ)[4] under this criterion, motivating the use of the alternate Akaike Information Criterion (AIC; Akaike, 1974). In the AIC, the test statistic is calculated as $-2\log l_T + 2p$ where p is the number of parameters used in the likelihood calculation for a tree T, l_T. An extra parameter is favored if it improves the likelihood by one log unit.

A third criterion, Bayesian Information Criterion (BIC), penalizes extra parameters more harshly with a term that depends on the data size n (Schwaz, 1978).

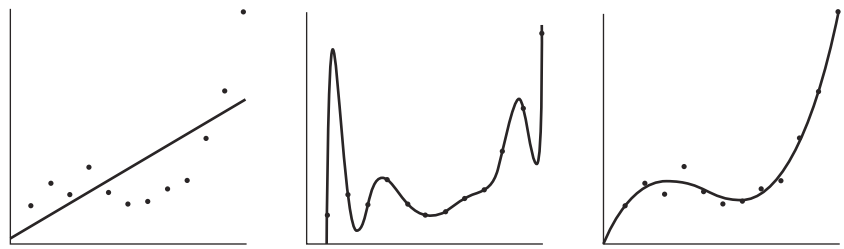


Figure 11.16: Data with various polynomial curves fitted to them.

[3]Even for JC69 there are other parameters in an analysis—one for each edge of the tree for example—but this is constant among models analyzing the same data set, hence plays no role in the marginal complexity of one model over another.

[4]The use of invariant sites simultaneously with Γ classes is problematic, since the parameters are not independent.

Model	*l*	K	AIC$_c$	Δ AIC$_c$	w	Cum(w)
TN93+I+Γ	5441.4600	78	11045.5888	0.0000	0.5221	0.5221
TIM+I+Γ	5441.3765	79	11047.5965	2.0077	0.1913	0.7134
HKY85+I+Γ	5443.6729	77	11047.8422	2.2534	0.1692	0.8826
K81uf+I+Γ	5443.5566	78	11049.7821	4.1934	0.0641	0.9468
GTR+I+Γ	5440.9150	81	11051.0301	5.4413	0.0344	0.9811
TVM+I+Γ	5442.7393	80	11052.4991	6.9103	0.0165	0.9976
TN93+Γ	5448.6792	77	11057.8549	12.2661	0.0011	0.9988
HKY85+Γ	5450.5068	76	11059.3402	13.7514	0.0005	0.9993
TIM+Γ	5448.6577	78	11059.9843	14.3955	0.0004	0.9997
K81uf+Γ	5450.4883	77	11061.4730	15.8843	0.0002	0.9999
GTR+Γ	5448.0298	80	11063.0802	17.4914	0.0001	1.0000
TVM+Γ	5449.6685	79	11064.1804	18.5917	0.0000	1.0000
TN93+I	5470.7568	77	11102.0102	56.4214	0.0000	1.0000
TIM+I	5470.7417	78	11104.1522	58.5635	0.0000	1.0000
GTR+I	5470.3452	80	11107.7110	62.1223	0.0000	1.0000
HKY85+I	5476.8496	76	11112.0257	66.4370	0.0000	1.0000
K81uf+I	5476.8208	77	11114.1381	68.5493	0.0000	1.0000
TVM+I	5476.1650	79	11117.1736	71.5849	0.0000	1.0000
F81+I+Γ	5769.1118	76	11696.5501	650.9614	0.0000	1.0000
F81+Γ	5782.0566	75	11720.2721	674.6834	0.0000	1.0000
F81+I	5807.4927	75	11771.1442	725.5554	0.0000	1.0000
GTR	5805.0576	79	11774.9588	729.3700	0.0000	1.0000
TVM	5808.4727	78	11779.6141	734.0254	0.0000	1.0000
TIM	5810.4102	77	11781.3168	735.7280	0.0000	1.0000
TN93	5813.4780	76	11785.2825	739.6938	0.0000	1.0000
K81uf	5813.5190	76	11785.3646	739.7758	0.0000	1.0000
HKY85	5816.5894	75	11789.3375	743.7488	0.0000	1.0000
SYM+I+Γ	5861.0859	78	11884.8407	839.2520	0.0000	1.0000
TVMef+I+Γ	5867.6128	77	11895.7221	850.1333	0.0000	1.0000
SYM+Γ	5876.7803	77	11914.0570	868.4683	0.0000	1.0000
TVMef+Γ	5884.4272	76	11927.1810	881.5922	0.0000	1.0000
TIMef+I+Γ	5885.0684	76	11928.4632	882.8745	0.0000	1.0000
K81+I+Γ	5893.7642	75	11943.6872	898.0984	0.0000	1.0000
TN93ef+I+Γ	5897.7529	75	11951.6647	906.0759	0.0000	1.0000
TIMef+Γ	5899.2588	75	11954.6764	909.0877	0.0000	1.0000
K80+I+Γ	5906.2329	74	11966.4593	920.8706	0.0000	1.0000
K81+Γ	5908.7876	74	11971.5687	925.9800	0.0000	1.0000
TN93ef+Γ	5911.5659	74	11977.1254	931.5366	0.0000	1.0000
SYM+Γ	5908.7021	77	11977.9008	932.3120	0.0000	1.0000
TVMef+I	5917.6128	76	11993.5521	947.9633	0.0000	1.0000
K80+Γ	5920.9038	73	11993.6382	948.0494	0.0000	1.0000
TIMef+I	5928.9629	75	12014.0846	968.4959	0.0000	1.0000
K81+I	5938.0137	74	12030.0209	984.4321	0.0000	1.0000
TN93ef+I	5940.7383	74	12035.4701	989.8813	0.0000	1.0000
K80+I	5949.5186	73	12050.8677	1005.2789	0.0000	1.0000
F81	6088.2227	74	12330.4388	1284.8501	0.0000	1.0000
JC69+I+Γ	6101.2656	73	12354.3618	1308.7730	0.0000	1.0000
JC69+Γ	6114.8408	72	12379.3515	1333.7628	0.0000	1.0000
JC69+I	6142.1719	72	12434.0137	1388.4249	0.0000	1.0000
SYM	6170.8916	76	12500.1097	1454.5209	0.0000	1.0000
TVMef	6190.3394	75	12536.8375	1491.2488	0.0000	1.0000
TIMef	6194.5806	74	12543.1547	1497.5659	0.0000	1.0000
TN93ef	6210.6353	73	12573.1011	1527.5123	0.0000	1.0000
K81	6214.1152	73	12580.0610	1534.4723	0.0000	1.0000
K80	6230.2100	72	12610.0898	1564.5011	0.0000	1.0000
JC69	6411.5161	71	12970.5438	1924.9551	0.0000	1.0000

Figure 11.17: Model test (Posada and Buckley, 2004) based on mitochondrial data of Sota and Vogler (2001). l is the log likelihood, K the number of parameters, AIC_l the Akaike Information Criterion, $DeltaAIC_l$ the difference in AIC_l with the next "best," w the Akaike weights, and $Cum(w)$ the cumulative Akaike weights.

$$\text{BIC} = -2 \log l_T + p \log n \tag{11.26}$$

These tests are implemented in Posada and Crandall (1998) and well summarized in Posada and Buckley (2004). An example of a test of a broad variety of models in an empirical context is given by Posada and Buckley (2004) in their reanalysis of Sota and Vogler (2001) (Fig. 11.17).

11.7 Exercises

1. What is the probability of transformation between the aligned sequences ACGT and AGCT under the JC69 model with time parameters $\mu t = \{0.1, 0.2, 0.5, 1.0\}$?

2. What is the probability of transformation between the aligned sequences $\frac{ACGT}{AGCT}$, $\frac{ACG-T}{A-GCT}$, and $\frac{A-CGT}{AGC-T}$ under a 5-state Neyman model with time parameters $\mu t = \{0.1, 0.2, 0.5, 1.0\}$?

3. What were the maximum likelihood estimators of the time parameter in the previous two exercises? If the four given time parameter values were the only ones possible, what would the integrated likelihoods be? What fraction of the integrated likelihoods were the maximum values?

4. Using a Neyman model for binary characters, what would the likelihoods be for the two cladograms in Fig. 11.18 where all time parameters (μt) were 0.1? 0.2?

5. Under the No-Common-Mechanism model, what are the likelihoods for the cladograms in exercise 4?

6. Two systematists argue the question "ML using No-Common-Mechanism and parsimony will yield the same tree for this data set," one taking the affirmative and one the negative, who is correct? Why?

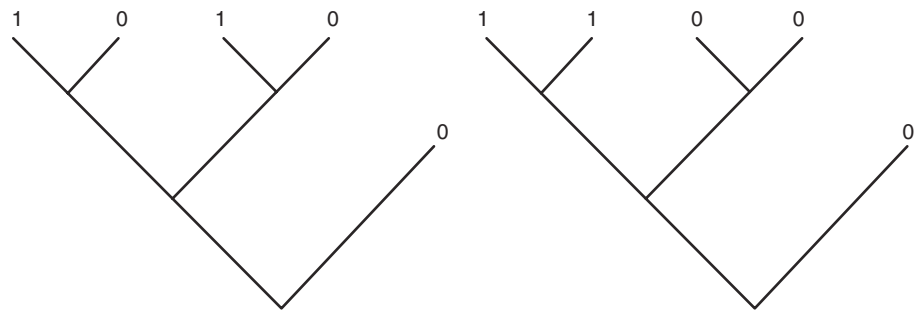


Figure 11.18: Example cladograms.

7. Using a Neyman model for sequence characters, and sequences ACGT and AGCT with time parameter $\mu t = 0.2$, determine the maximum likelihood alignment of the two sequences. What are the values of the "total" and "dominant" likelihoods? What fraction of the total likelihood is the dominant? Give an example of a non-dominant likelihood alignment (and its likelihood) included in the total likelihood calculation.