Reconstructing phylogenies from nucleotide pattern probabilities: A survey and some new results

Mike Steel\textsuperscript{a,\,*\,1} Michael D. Hendy\textsuperscript{b,\,2} David Penny\textsuperscript{c,\,3}

\textsuperscript{a} Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand
\textsuperscript{b} Mathematics Department, Massey University, Palmerston North, New Zealand
\textsuperscript{c} Molecular Genetics Unit, Massey University, Palmerston North, New Zealand

Received 27 May 1997; received in revised form 10 November 1997; accepted 15 April 1998

Abstract

The variations between homologous nucleotide sequences representative of various species are, in part, a consequence of the evolutionary history of these species. Determining the evolutionary tree from patterns in the sequences depends on inverting the stochastic processes governing the substitutions from their ancestral sequence. We present a number of recent (and some new) results which allow for a tree to be reconstructed from the expected frequencies of patterns in its leaf colorations generated under various Markov models. We summarize recent work using Hadamard conjugation, which provides an analytic relation between the parameters of Kimura's 3ST model on a phylogenetic tree and the sequence patterns produced. We give two applications of the theory by describing new properties of the popular "maximum parsimony" method for tree reconstruction. © 1998 Elsevier Science B.V. All rights reserved.

Keywords: Trees; DNA sequences; Markov processes; Maximum parsimony; Hadamard matrices; Site substitution models

1. Introduction

A fundamental problem in biological classification is the following: how can the large and rapidly expanding array of DNA and RNA sequences be best exploited to provide an accurate picture of how species evolved from common ancestors? It is increasingly recognized that approaches to this question should be statistically based [34]. This requires the underlying sequence evolution to be modelled stochastically, and a variety of models have been proposed. We first describe a number of classes of

\textsuperscript{\,*} Corresponding author. Tel.: 64-3-3667001 (ext. 7688); fax: 64-3-3642 587; e-mail: m-steel@math.canterbury.ac.nz.
\textsuperscript{1} Supported by New Zealand Lotteries Commission and the Alexander von Humboldt-Stiftung.
\textsuperscript{2} Partially supported by the DAAD.
\textsuperscript{3} Partially supported by the VW foundation.

0166-218X/98/$19.00 © 1998 Elsevier Science B.V. All rights reserved.
PHS S0166-218X(98)00080-8
such models. We then discuss the fundamental and biologically important "inversion" problem of reconstructing trees uniquely, given only the expected frequencies of their induced leaf colorations (patterns). This provides the mathematical basis for statistical approaches to phylogeny reconstruction, where the frequencies of patterns in finite length sequences are approximations to these expected frequencies.

Next we review two important, classical characterizations of phylogenetic trees — as set systems, and as distance functions satisfying a "four point condition". In Section 2.1 we describe how the second of these characterizations allows for the inversion problem to be solved with very few assumptions regarding the associated transition matrices.

We then consider more specific models in which progressively more structure is imposed on the model: from insisting that the transition matrices belong to some prescribed semigroup, and special cases of such models — Stationary models and Group-based models for which we present new results for these models in Theorems 4 and 5. Group-based models include the symmetric model for 2 colors (based on the cyclic group $C_2$) and an extension of the biologically relevant 3-parameter model due to Kimura [32] which corresponds to the group $C_2 \times C_2$. Recently, it has been shown that both these models (and others based on abelian groups, particularly elementary abelian groups) result in a particularly nice and fully invertible relationship between a tree and the frequencies of the patterns it induces (see [25, 44, 50]). This is summarized in Section 2.2.3 using the characterization of a phylogenetic tree as a set system from Section 1.2.1. In particular, for the Kimura 3ST model [32], we present a self-contained and transparent proof of the main inversion theorem from Steel et al. [44], which complements the more abstract approach to group-based models (based on discrete Fourier analysis) adopted by Székely et al. [44, 50]. We also summarize in Section 2.3 some recent extensions that allow for a distribution of rates across sites.

In Section 3 we present two new applications of the theory to analyze the popular tree building method based on "maximum parsimony": (1) we show that this method is statistically consistent on four species, under Kimura's 3ST model, with a molecular clock, and (2), under the symmetric 2-color model we prove the "Bealej Theorem" which bounds the expected number of sites requiring 2, 3, ..., substitutions on the true tree in terms of the expected number requiring 0 and 1 substitution (regardless of the parameters on the underlying tree); an application of this theorem to biological data has already appeared in [35].

1.1. General formulation

In this section we provide the framework from which we formulate the inversion problem and detail some assumptions necessary for this inversion.

1.1.1. Randomly coloring phylogenetic trees

Evolutionary relationships are generally represented by a phylogenetic tree, $T$, that is, a tree whose leaves are labelled (bijectively) by a set $S$ of species and whose remaining vertices are unlabelled and of degree at least 3. When all the non-leaf vertices have
degree 3 the tree is said to be fully resolved. If we take a phylogenetic tree $T$ and either distinguish a non-leaf vertex by labelling it $\rho$, or bisect an edge of $T$ and label the newly created degree 2 vertex $\rho$, the resulting tree, denoted $T^{+\rho}$, is called a rooted phylogenetic tree. In taxonomy, the leaves of $T$ and $T^{+\rho}$ generally represent extant species, the remaining vertices represent ancestral species. The root vertex $\rho$ in $T^{+\rho}$ represents the most recent common ancestor of the species set $S$.

We represent the assignment of characters of biological interest as a coloring of the vertices of $T^{+\rho}$. Direct the edges of $T^{+\rho}$ away from $\rho$, and for each edge $e$, we write $e$ as the ordered pair $(u,v)$ if $u$ lies between $v$ and $\rho$. Consider the following probability distribution on the set of leaf-colorations of $T$ by elements of a set $C$ of $c$ colors. First, assign a color $z \in C$ to the root vertex $\rho$ with probability $\pi_z(\rho)$. Then, randomly color the remaining vertices of $T^{+\rho}$ recursively, from the root towards the leaves, as follows: if $e = (u,v)$ has vertex $u$ assigned a color, say $z$, and $v$ is yet to be colored, then assign a random color $\beta$ to $v$ with probability $p_{\beta}(z,\beta)$. Eventually all the vertices of $T^{+\rho}$, including the leaves of $T^{+\rho}$, will be colored, and each such total coloration (coloration of all the vertices of $T^{+\rho}$) $\tilde{\gamma}$ produced in this way will have a certain probability. Now, suppose we are given a coloration $\chi$ of $S$ by $C$ - we call this a pattern on $S$. If we regard $\tilde{S}$ as the set of leaves of $T^{+\rho}$ then $\chi$ has an induced marginal probability, equal to the sum of the probabilities of that subset of the total colorations which extend $\chi$.

We denote by $f_\chi$, the probability of generating pattern $\chi$, so that

$$f_\chi = \sum_{\tilde{\gamma}} \pi_{\tilde{\gamma}}(\rho) \prod_{e=(u,v)} p_{\beta}(\tilde{\gamma}(u),\tilde{\gamma}(v)), \quad (1)$$

where the summation is over all total colorations $\tilde{\gamma}$ which extend $\chi$ and the product is over all edges of $T^{+\rho}$. Note that if $T^{+\rho}$ has $\omega$ non-leaf vertices there will be $e^{\omega}$ such extensions.

For $e=(u,v)$, an edge of $T^{+\rho}$, we will let $M(e)$ denote throughout the transition matrix $M(e) = [p_{\beta}(z,\beta)]$. Thus, if we order $C$ as $(z_1,\ldots,z_c)$, the $rs$ entry of $M(e)$ is the conditional probability that $\tilde{\gamma}(v) = z_r$, given that $\tilde{\gamma}(u) = z_s$. Consequently, each row of $M(e)$ sums to 1. As a simple example, consider the tree in Fig. 1, together with the indicated $2 \times 2$ transition matrices and the root distribution $\pi(\rho) = (\pi_1,\pi_2)$. The probability of the pattern $\chi(1) = \chi(2) = z_1, \chi(3) = z_2$, is

$$f_\chi = \pi_1[(1-p_1)(1-p_2)(1-p_3)p_4 + (1-p_1)p_2q_3(1-q_4)]$$

$$+ \pi_2[q_1q_2(1-p_3)p_4 + q_1(1-q_2)q_3(1-q_4)].$$

In our recursive description above of how to generate random patterns based on $\pi(\rho)$ and $\{M(e)\}$, we have tacitly assumed that each new coloring of a vertex is dependent only on the color of its immediate ancestor. In the interests of precision we now make explicit this assumption. Let $\prec$ be a total ordering on the vertices of $T$ that respects descendency from the root. so that if $e = (u,v)$, then $u \prec v$ (hence for example $\prec$ may
be induced by time). Then in order for (1) to hold, we need only assume the following equality of conditional probabilities for each edge \( e = (u, v) \) of \( T^{+\rho} \),

\[
(A1) \quad \mathbb{P} \left[ \tilde{X}(v) = \alpha \bigg\vert \bigwedge_{w \leftarrow v} \tilde{X}(w) \right] = \mathbb{P}[\tilde{X}(v) = \alpha \bigg\vert \tilde{X}(u)].
\]

Informally, (A1) states that given the state at vertex \( u \), the state assigned to vertex \( v \) is conditionally independent of the states at all other "earlier" vertices. (A1) implies Eq. (1) by the well-known identity in probability theory, for a family of events \( A_1, A_2, \ldots \),

\[
\mathbb{P} \left[ \bigcap_i A_i \right] = \mathbb{P}[A_1] \mathbb{P}[A_2 \mid A_1] \mathbb{P}[A_3 \mid A_1 \land A_2] \cdots
\]

If the tree \( T^{+\rho} \) consisted of a path from \( \rho \) to a single leaf, then (A1) would be precisely the definition of a non-homogeneous Markov chain (see [30, chapter 7]). Thus, (A1) defines what one might call a "nonhomogeneous Markov tree".

1.1.2. Inversion

A fundamental issue for phylogenetic methodology is the inverse problem, of finding \( T^{+\rho} \) and \( \{M(e)\} \), or relevant information about these matrices, given just the probabilities of the various patterns on \( S \) together with certain restrictions on \( \{M(e)\} \). If \( \{M(e)\} \) is not required exactly, it may still be desirable to determine, or at least place bounds on the "edge lengths" – that is the expected number of changes of color on each edge under the assumption that the transition matrix for that edge is the result of a continuous-time Markov process (see Remark 2.2.1, below). For stationary models (discussed in Section 2.2.1) these lengths are proportional to time, so that their determination allows for the temporal dating of different evolutionary episodes. As an intermediate step, it would be desirable to at least be able to order the vertices of \( T^{+\rho} \) consistently with the temporal order of the evolutionary events that such vertices represent (namely the creation of new species from an ancestral species).
Actually, as we shall see, the position of the root in $T^{+\rho}$ cannot be uniquely established without invoking additional assumptions — in taxonomy the inclusion of an additional outgroup species, or the imposition of a hypothesis such as the molecular clock (discussed below) are used to estimate the position of the root. Thus, a more reasonable goal is the following:

- **Tree reconstruction problem:** Given $f = \{f_x\}$, or some knowledge of its distribution, find $T$, and information about $\{M(e)\}$.

In taxonomic applications, $f_x$ is usually estimated as the observed proportion of sites in a collection of aligned sequences which correspond to $\chi$. Provided the sites in the sequence have evolved identically and independently (i.i.d.) according to the above model, these estimates will tend, with probability 1, to the true probability value as the length of the sequences increases. In taxonomy, with sequences of finite length, statistical methods must be appended to a solution of the inversion problem in order to determine confidence limits for reconstructed trees (we do not consider these here, see, for instance, [51]). Also in taxonomy the assumption that the sites have an identical distribution satisfying (A1) is often violated, however we describe how, for certain models that allow the rate of evolution to vary across sites, the inversion problem can still be solved.

Note that restrictions must be placed on $\{M(e)\}$ for $T$ to be uniquely described by the $f_x$'s. For example, with only two colors, putting $p_e(\alpha, \beta) = 0$ for all edges $e$, we see that all phylogenetic trees induce exactly the same distribution on the set of leaf bicolorations (namely the degenerate distribution which colors all the leaves $\alpha$ with probability $\pi_\alpha(\rho)$). Similarly, setting $p_e(\alpha, \beta) = 0.5$ for all $\alpha, \beta$, we obtain the uniform distribution on the $\chi$'s.

A further technical point concerns the occasional practice in taxonomy of grouping the four nucleotide bases into the two purine bases and the two pyrimidine bases, thereby replacing 4-colorations of the vertices of $T^{+\rho}$ by 2-colorations; although assumption (A1) may apply for four bases, (A1) may fail when the four colors are grouped into pairs.

1.2. Representations of phylogenetic trees

In this section we review two fundamental theorems concerning phylogenetic trees, both of which provide neat existence and uniqueness results for a tree in terms of an induced structure, and are central to later sections.

1.2.1. A phylogenetic tree as a system of splits

Normally a phylogenetic tree is thought of as a graph. However, there is a natural way to represent an (unrooted) phylogenetic tree on a leaf set $S$ as a collection of subsets of $S$, and this representation is an essential aspect of inversion formulae discussed later. If we take a phylogenetic tree, $T$, with leaf set $S$ and we delete an edge of $T$, this disconnects $T$ into two components and thereby partitions $S$ into a pair of subsets; this pair is frequently referred to as a *split*. If we distinguish one element $R$ of $S$, one
of the two subsets in a split will not contain $R$. We select this subset to identify the split. The collection of these split identifying subsets for all the edges of the tree $T$ is a collection, $\sigma = \sigma(T)$, of nonempty subsets of $S' = S \setminus \{R\}$ which have the following two properties:

(i) \( S' \in \sigma \) and \( \{i\} \in \sigma, \forall i \in S' \),

(ii) if \( \beta, \beta^* \in \sigma \) then \( \beta \cap \beta^* \in \{\beta, \beta^*, \emptyset\} \).

Condition (ii) is often expressed by saying that $\beta$ and $\beta^*$ are *compatible*. $\sigma(T)$ has at most $2|S| - 3$ sets, and this upper bound is achieved precisely if $T$ is fully resolved [9]. For example, for the fully resolved tree $T_1$, in Fig. 2, taking $R = 4$, we have

$$\sigma(T) = \{\{1\}, \{2\}, \{3\}, \{1, 2, 3\}, \{1, 2\}\}.$$

Buneman [9] established the fundamental converse result:

**Theorem 1** (Buneman [9]). Any collection $\sigma$, of non-empty subsets of $S'$ which satisfy (i) and (ii) corresponds to $\sigma(T)$ for a unique unrooted phylogenetic tree $T$ on $S$. Furthermore, this tree can be recovered from $\sigma$ in polynomial time.

Methods for reconstructing $T$ from $\sigma$ include Meacham’s “Tree popping” method (see [2]), or Gusfield’s linear-time method [24]. More generally, Buneman [9] described a natural association of a graph to any collection $\sigma$ of subsets of $S'$, and showed that this graph is a tree $T$ precisely if the sets in $\sigma$ are all pairwise compatible (in which case $\sigma = \sigma(T)$). For further details the interested reader should consult [8]. Unfortunately, for this construction the number of vertices in the graph can grow exponentially with $n = |S|$. A preferable graphical representation of $\sigma$ – which extends to positively weighted splits – is provided by the recently developed split decomposition method [3]. In this representation, “weakly compatible” sets of positively weighted splits induce an edge-weighted graph with a small (order $n^2$) number of vertices, and this graph is a tree exactly when the splits are pairwise compatible.

### 1.2.2. A phylogenetic tree as a distance function

A *distance function* on $S$ is a map $d : S \times S \to \mathbb{R}^{\geq 0}$ (the non-negative real numbers) which is symmetric (i.e., $d(x, y) = d(y, x)$ for all $x, y \in S$) and for which $d(x, x) = 0$ for all $x \in S$. A (rooted or unrooted) phylogenetic tree $T$ whose edges are weighted
according to non-negative real-valued function $\lambda$, induces a distance function $d = d(T, \lambda)$ on the leaf set $S$ by simply letting $d_{ij}$ be the sum of $\lambda(e)$ over all edges $e$ on the path in $T$ connecting $i$ and $j$. That is,

$$d_{ij} = d_{ij}(T, \lambda) := \sum_{e \in P(T; i, j)} \lambda(e),$$

where $P(T; i, j)$ is the path in $T$ connecting leaves $i$ and $j$. If a distance function $d$ on $S$ can be expressed in this way then $d$ is said to be additive on $T$, and $\lambda$ is said to realise $d$ on $T$.

Such a $d$ not only satisfies the triangle inequality, it also satisfies a stronger “four point” condition:

- For any four leaves $i,j,k,l$, (not necessarily distinct),

$$d_{ij} + d_{kl} \leq \max\{d_{ik} + d_{jl}, d_{il} + d_{jk}\}. \tag{2}$$

This condition is equivalent to the following: of the three pairwise sums ($d_{ij} + d_{jk}, d_{ik} + d_{jl}$ and $d_{il} + d_{jk}$) two sums are equal and they are at least as large as the other sum. Several independent proofs (see [1] for references) have been given of the following fundamental result:

**Theorem 2** (Buneman [9]). (1) A distance function $d$ on $S$ is additive on some tree if and only if $d$ satisfies the four point condition (Eq. (2)).

(2) If $d$ is additive on some tree, then there exists only one pair $(T, \lambda)$, where $T$ is an unrooted phylogenetic tree, and $\lambda$ is a non-negative edge weighting of $T$, with $\lambda(e) > 0$ if $e$ is not incident with a leaf, and such that $\lambda$ realizes $d$ on $T$. Both $T$ and $\lambda$ can be constructed from $d$ efficiently (i.e. in polynomial time).

Both parts of this theorem have natural analogues when $\lambda$ is allowed to be any real-valued function defined on the edges of $T$, or, more generally, where $\lambda$ takes values in a suitably structured abelian monoid (for details, see [4]). A useful connection between these two representations of trees – as splits and as a distance function - is given by the “isolation index” of a split – for details and extensions the interested reader is referred to [3].

2. Varieties of models and their inversion

We now describe some of the types of models that arise when additional assumptions are added to the Markov property (A1). The most general of these seeks only to avoid the null and random effects, while more structured models require some semigroup structure in the stochastic process. In all cases we show how the expected frequencies of patterns on $S$ can be used to identify the generating tree uniquely and sometimes additional information, such as the “lengths” of the edges.
2.1. The general model

As mentioned earlier, even with two colors, and \( M(e) \) symmetric for all edges \( e \), it is not always possible to recover the unrooted tree \( T \), because if we set both off-diagonal entries in all the matrices \( M(e) \) either to 0 or to 0.5, then every rooted phylogenetic tree \( T^{+\rho} \) induces exactly the same distribution on the set of leaf bicolorations. Note that in these cases \( \det(M(e)) = 1 \) and 0 respectively, where \( \det(M) \) denotes throughout the determinant of matrix \( M \). In the general model, in addition to (A1), we therefore make the following mild (and biologically reasonable) assumption:

(A2) For all edges \( e \) of \( T^{+\rho} \), \( \det(M(e)) \neq 0, \pm 1; \pi_z(\rho) \neq 0 \), for all colors \( z \in C \).

It is easily shown that \( \det(M(e)) = 1(-1) \Leftrightarrow M(e) \) is an even (odd) permutation matrix. The general model therefore allows \( c^2 - c \) parameters in each transition matrix (subject only to non-negativity, the \( c \) linear stochastic equations, and the inequalities of condition (A2)) and so is of the type of model discussed in [7, 13] and (for 2-color characters) [41].

Theorem 3 (below) shows that conditions (A1) and (A2) are sufficient for the \( f_{ij} \)'s to uniquely determine \( T \). However as shown in [50], the root \( \rho \) cannot be located on \( T \) under assumption (A2) alone.

Note that (A2) does not require \( M(e) \) to be diagonalizable, nor to have all its eigenvalues real. Also, the general model does not make any assumption about the actual process occurring on an edge which produces net random transitions of colors between its ends, in particular it does not assume any sort of fixed continuous-time process, let alone a “rate” matrix constant across edges of the tree (as in the stationary models discussed below). Since we do not make any further assumption about the root distribution \( \pi(\rho) \) or the structure on the family of transition matrices, apart from those properties prescribed by (A2), the model is valid for a much wider class of models than is usually considered in molecular taxonomy (see [42]). We now describe an analytical result which shows that \( T \) can be easily and quickly reconstructed from the \( f_{ij} \)'s in the general (nonsymmetric) case, with any number of colors, under assumptions (A1) and (A2).

Let \( C = \{x_1, \ldots, x_c\} \) be the set of \( c \) colors, and for any vertex \( u \) of \( T^{+\rho} \) let \( \pi_k(u) \) be the probability that vertex \( u \) is assigned the color \( x_k \). (By (A1) this will be a function of \( \pi(\rho) \) and the transition matrices on the path from \( \rho \) to \( u \).) Let \( \Pi(u) = \text{diag}[\pi_1(u), \ldots, \pi_c(u)] \) (the diagonal matrix with \( \pi_k(u) \) as its \( (k, k) \) entry) and for leaves \( i, j \) of \( T^{+\rho} \), let \( F_{ij} = [f_{ij}(k, l)] \) be the \( c \times c \) “divergence” matrix with \( (k, l) \) entry \( f_{ij}(k, l) \), the probability that leaf \( i \) is colored \( x_k \) and leaf \( j \) is colored \( x_l \).

**Theorem 3.** Under the general model, with underlying generating tree \( T^{+\rho} \),

\[
\phi_{ij} := -\frac{\ln[|\det(F_{ij})|] + 0.5(\ln[\det(\Pi(i)\Pi(j))])}{c},
\]

is a well-defined distance function, which is additive on (and hence defines) \( T \).
Thus, each phylogenetic tree $T$ (without specifying the placement of its root) is uniquely defined by the collection of probabilities of the patterns it induces under assumptions (A1) and (A2), and it can be reconstructed from the $f_{ij}$ values in polynomial time. The condition $\pi_s(\rho) \neq 0$ can be relaxed for tree recovery, although in that case the function $\phi_{ij}$ of (3) is infinite for all pairs of vertices $i, j$ separated by $\rho$.

Variations on the Theorem 3 are due, independently, to Steel [43], Lake [33] and Chang and Hartigan [13]. Lake [33] refers to $\phi_{ij}$ as “Paralinear distance”. Barry and Hartigan [6] defined a similar, but different measure, based on the logarithm of the determinant of the conditional (rather than joint) probability distribution on the colors of leaves $i$ and $j$. Consequently, their measure $\phi'_{ij}$ does not have the tree-like property described for $\phi_{ij}$ in the following theorem – in fact, as they point out, it is not even symmetric with respect to $i$ and $j$, whereas, from (3), $\phi_{ij} = \phi_{ji}$. However, $\phi_{ij} = \frac{1}{2}(\phi'_{ij} + \phi'_{ji})$, and in [13], Theorem 3 is stated without proof (later provided in [12]). Similar ideas have also been developed, independently, in [53].

With a finite sample of sites generated by the general model we can only estimate the $f_{ij}$ values, which suggests the following procedure (provided the number of sites is large):

- Step 1. For each pair of leaves $i, j$, and each $k, l = 1, \ldots, c$, estimate $f_{ij}(k, l)$ by setting it equal to the proportion of sites in which $i$ and $j$ are colored $x_k$ and $x_l$, respectively.
- Step 2. Using (3), calculate $\phi_{ij}$ for each pair $i, j$ using the entries from step 1.
- Step 3. Use a suitable distance-based tree reconstruction method, using the $\phi_{ij}$ values from step 2.

By a “suitable” method in step 3 it is desirable to use a method which can be implemented in reasonable time, even for large values of $n = |S|$, and which, as a map from distance functions on $S$ onto the subspace of additive distance functions on $S$ has the properties that it:

(i) fixes every additive distance function,
(ii) is continuous in a neighborhood of each additive distance function.

There are many such methods, one of the earliest being the Buneman retraction [9], which has the stronger property of being continuous on the entire space of distance functions (unlike other methods, such as neighbor-joining), see [37] for a proof.

Under these conditions, and provided the sites evolve i.i.d. such a method will be statistically consistent in the following sense: as the number of sites grows, the reconstructed tree will (with probability tending to 1) be the true tree with, perhaps, some additional (short) edges, but the maximum length of these “phantom” edges will go to zero (with probability tending to 1) as the number of sites tends to infinity. In case the true tree is fully resolved, then these phantom edges eventually disappear entirely and the reconstructed phylogenetic tree will actually equal the true tree given sufficient sites (see [12] for a discussion of this issue, in relation to maximum likelihood).

An application of the above procedure to biological data is given in [36], where it is also extended with the deletion of a proportion of constant (uniformly colored) sites,
under the assumption that this proportion represents the number of invariant sites, with
the remaining sites evolving i.i.d.

Theorem 3 can also be used to show that the maximum (average) likelihood method
described by Barry and Hartigan [7] will identify uniquely the correct tree given suffi-
cient data, under the general model, and assuming the underlying tree is fully resolved
(for a proof see [12]). A maximum likelihood approach may be preferable, particularly
from the perspective of statistical efficiency to the above procedure if only a moderate
number of sites is available. Indeed the above procedure will not work if any of the
matrices \( F_{ij} \) is singular, which can occur with a small number of sites. In any case it
is useful to estimate the variance of the \( \phi_{ij} \) values, and also correct statistical bias (see
[5, 22, 36]).

Once \( T \) has been reconstructed, it is natural to ask if the transition matrices \( M(e) \)
and the root distribution \( \pi(p) \) can also be recovered from \( f = [f_i] \). Of course these
parameters apply to \( T^{+ \rho} \), which differs from \( T \) if \( \rho \) has degree 2, but in this case, by
re-rooting \( T \) on any other vertex \( \rho' \), it is possible to assign a distribution \( \pi(\rho') \) of colors
to this vertex, and transition matrices \( M'(e) \) to the edges of \( T^{+ \rho'} \) in such a way that
the induced distribution on patterns is precisely \( f \) (and furthermore, \( \pi(\rho') \) and \( \{M'(e)\} \)
satisfy (A2) – for details see [47]), and so it is not possible to recover the position
of the root just from \( f \). Furthermore, the distribution of patterns on pairs of leaves
does not suffice to determine the parameters \( \{(M'(e)), \pi(\rho')\} \), however under certain
restrictions on the underlying parameters, the distribution on \textit{triples} of leaves does.
These last two results are due to Chang [12] (who extended earlier results confined to
two-color characters, by Pearl and Tarsi [41]). We now describe additional constraints
which are frequently imposed upon the family \( \{M(e)\} \), and the implications these have
for the reconstruction problem.

### 2.2. Semigroup models

Semigroup models assume that the transition matrices \( M(e) \) all belong to some
prescribed semigroup. An example is the general model in which the transition matrices
satisfy condition (A2) (\( \det(M(e)) \neq 0, \pm 1 \)). A (commutative) semigroup of transition
matrices arising with \( c = 2 \) colors is the family:

\[
M(e) = \begin{bmatrix}
1 - p(e) & p(e) \\
xp(e) & 1 - xp(e)
\end{bmatrix}
\]

where \( x > 0 \) is independent of \( e \), and \( 1 - p(e)(1 + x) > 0 \) (this last constraint is imposed
in order for \( \det(M(e)) > 0 \)). If \( x = 1 \), we obtain the \textit{2-color Neyman model} ([39]; see
also [10, 17]).

A number of biologically relevant semigroups for four color models have been stud-
ied, for example, the six-parameter unbalanced transversion model, see Nguyen and
Speed [40]. We now consider two important subclasses of semigroup models.
2.2.1. Stationary models

These are based on (A1) and three further assumptions:

(i) Color changes on edges are described by a continuous time Markov process.

(ii) The associated intensity matrix \( R \) is the same for all edges of the tree.

(iii) The distribution of colors at the root of the tree is the equilibrium distribution.

A number of stationary models of relevance to taxonomy have been described and studied by Rodriguez et al. [42]. Note that conditions (i)–(iii) can be restated as follows:

\[
M(e) = \exp(R^{\alpha_e}) = I + \sum_{k > 0} \frac{R^k}{k!} \alpha_e^k,
\]

(4)

\[
\pi R = 0,
\]

(5)

where \( \alpha_e > 0 \) is a parameter associated with edge \( e \) and where \( \pi = [\pi_1, \ldots, \pi_c] \), (and where \( \pi_i = \pi_i(\rho) \) for \( i = 1, \ldots, c \)). The matrix \( R \) is often called the (substitution) rate matrix. A further condition which is sometimes imposed is the molecular clock hypothesis which states that the sum of the \( \alpha_e \)'s on the path in \( T \) from \( \rho \) to any leaf \( x \) is the same for all \( x \), and so the \( \alpha_e \)'s are proportional to time (we do not assume this here, except in Section 3.1). Condition (5) asserts that the colors at the root are in equilibrium, thus the probability distribution of colors at any individual vertex of the tree is also \( \pi \). Note that the rows of \( R \) sum to 0, and since \( R_{ij} > 0 \) for \( i \neq j \) it follows that \( \text{tr}(R) < 0 \), where “\( \text{tr} \)” refers to the matrix trace function. The condition \( \text{tr}(R) < 0 \) together with Jacobi’s identity (see [21]):

\[
\det(\exp(M)) = \exp(\text{tr}(M))
\]

(6)

applied to \( M = R^{\alpha_e} \), shows that stationary models satisfy not only the first part of condition (A2) but the stronger constraint,

\[(A2') \quad 1 > \det(M(e)) > 0 \quad \text{for all edges } e \text{ of } T.\]

Stationary processes also lead to transition matrices which form a semigroup, by virtue of the identity:

\[
\exp(R^{\alpha_e}) \exp(R^{\alpha_e'}) = \exp(R^{\alpha_e + \alpha_e'}).
\]

An important class of stationary models are the reversible models, which assume in addition that \( \Pi R \) is symmetric, where \( \Pi \) is the diagonal matrix \( \text{diag} [\pi_1, \ldots, \pi_c] \). This condition implies that the Markov chain with transition matrix \( M(e) \) is reversible (see [42]). Examples of reversible models are the symmetric models for which \( R = R^t \) (which implies that \( \pi_1 = \cdots = \pi_c = 1/c \); the converse is true only for two colors). More generally, the matrices corresponding to reversible models are precisely the matrices \( R \) which can be obtained by multiplying the row \( i \) \( (i = 1, \ldots, c) \) of a symmetric rate matrix \( Q \) by \( x_i > 0 \). Thus, for four colors, each reversible model is defined by 9 free parameters (or 6 if we specify \( \pi \)), and with only two colors every stationary model is
reversible. In the latter (2-color) case the set of transition matrices forms the semigroup described at the beginning of Section 3. If we write the corresponding rate matrix as:

$$R = (1 + x)^{-1} \begin{bmatrix} -1 & 1 \\ x & -x \end{bmatrix}, \quad x > 0,$$

(so that $\pi = [x/(1 + x), 1/(1 + x)]$ and $R^2 = -R$), then $M(e) = I + (1 - \exp(-\lambda_e))R$.

2.2.1.1. Dissimilarity. A common measure of the difference between two species $i$ and $j$ is the proportion of sites in a collection of aligned sequences at which the two species differ (proportional to the Hamming distance between the sequences). Let $p_{ij}$ denote the probability that leaves $i$ and $j$ are differently colored (which can be estimated from the sequences by the Hamming distance). Note that $p_{ij} = 1 - \text{tr}(F_{ij})$.

In Theorem 3 we derived from $F_{ij}$ a measure which was “tree-like” (i.e., satisfied the four point condition); in Theorem 4 we show that such a tree-like measure can be calculated just from $p_{ij}$. This is relevant in biology where $p_{ij}$ is sometimes estimated from dissimilarity values and where the full divergence matrix $F_{ij}$ may not be available.

Of particular interest is the relationship between $p_{ij}$ and the expected number of color changes (“substitutions”) occurring on the path joining leaves $i$ and $j$ for a stationary model. We denote this quantity by $\delta_{ij}$. Clearly,

$$\delta_{ij} = \sum_{e \in P(T : i, j)} \delta_e,$$

(7)

where $\delta_e$ is the expected number of color changes occurring on edge $e$.

Thus, $\delta_{ij}$ (but not $p_{ij}$) satisfies the four point condition described in Eq. (2), so that determining the $\delta_{ij}$ values allows for $T$ to be reconstructed, along with its “true” edge lengths (in terms of the expected number of substitutions). For a stationary model, we have, from [6] or [42]:

$$\delta_e = \left( \sum_{x \in C} -R_{xx} \pi_x \right) \lambda_e = -\text{tr}(PR) \lambda_e.$$  

(8)

The main result from Rodriguez et al. [42] is that for a certain class of stationary models (including all reversible models and all models for which $\pi$ is the uniform distribution), $\delta_{ij}$ can be calculated from the divergence matrix $F_{ij}$ (discussed in Theorem 3) and $\Pi = \text{diag} [\pi_1, \ldots, \pi_c]$ (which can in turn be estimated, for stationary models, from $F_{ij}$).

Their result states that

$$\delta_{ij} = -\text{tr}(\Pi \ln(\Pi^{-1}F_{ij}))$$

where for a matrix $M$, $\ln[M] = -\sum_{k > 0} (I - M)^k/k$, provided this sum converges. (Actually, Rodriguez et al. [42] assume a molecular clock, though their proof can easily be modified so as to apply without this assumption.)

In the special case where the root distribution $\pi$ is the uniform distribution, $\pi_k = 1/c$, then it is easy to show that $\delta_{ij} = \phi_{ij}$, where $\phi_{ij}$ is the additive quantity described in Theorem 3 (see [23]). Thus, in this special case, not only can $T$ be found, but in
addition, for each edge, the expected number of substitutions (the “edge length”) can be found. We now describe an invertible relationship between $\delta_{ij}$ and $p_{ij}$.

**Theorem 4.** For a stationary reversible model,

\[
p_{ij} = 1 - \text{tr} \left( \Pi \exp \left( \frac{-\delta_{ij}}{\text{tr} (\Pi R)^{-1}} R \right) \right).
\]

(9)

(2) Eq. (9) is invertible.

(3) If the reversible stationary model applies to just two colors, $\alpha, \beta$, then letting $\gamma = 2\pi_x\pi_\beta$, we have

\[
\delta_{ij} = -\gamma \ln \left( 1 - \frac{p_{ij}}{\gamma} \right).
\]

**Proof.** (1) We have $F_{ij} = M_i^j \Pi M_j$, where $\Pi = \text{diag} [\pi_1, \ldots, \pi_c]$, and for $x = i, j$,

\[
M_x = \prod_{e \in P(T^+; u, x)} \exp (R \lambda_e) = \exp (R \tilde{\lambda}_x),
\]

where

\[
\tilde{\lambda}_x = \sum_{e \in P(T^-; t, x)} \lambda_e
\]

and where $v$ denote the most recent common ancestor of $i$ and $j$ (the last vertex common to the paths in $T^+$ from $\rho$ to $i$ and to $j$). Now, since $R^t \Pi = \Pi R$, we have $M_i^j \Pi = \Pi M_j$ so that,

\[
1 - p_{ij} = \text{tr} (\Pi M_i^j) = \text{tr} (\Pi \exp (R \lambda_i + \lambda_j)) = \text{tr} \left( \Pi \exp \left( \frac{\delta_{ij}}{-\text{tr} (\Pi R)^{-1}} R \right) \right),
\]

from (7) and (8), giving (9).

(2) By the spectral theory for reversible Markov chains [38, pp. 32–34], one can write

\[
p_{ij} = 1 - \sum_{k=1}^{r} \alpha_k e^{-\beta_k \delta_{ij}},
\]

where $\alpha_k \geq 0$ and $\sum_{k} \alpha_k = 1$, and $\beta_k \geq 0$ with at least one $\beta_j > 0$ (provided $R \neq 0$). Thus $p_{ij}$ is a strictly monotone increasing function of $\delta_{ij}$ and so is invertible.

(3) Since there are just two colors, we can (as described above) scale $R$ so that $R^2 = -R$, and hence $\exp (\lambda R) = I + (1 - \exp (-\lambda)) R$, which makes the inversion of (9) straightforward. □

**Remark.** Suppose that for an edge $e = (u, v)$ of $T^{+\rho}, M(e)$ is described by a continuous-time Markov process (so that we could write $M(e) = \exp (R \lambda_e)$) but that the distribution $\pi = \pi(u)$ of colors at $u$ is not necessarily the equilibrium distribution (so that $\pi R \neq 0$).
In this case the expected number of substitutions on edge $e$ is $\delta_e = -\frac{1}{|V|} \ln(M(e))$ with $V = \text{diag}(v_1, \ldots, v_c)$, where $v_i = \int_0^1 \phi_i(t) \, dt$, and for which $[\phi_1(t), \ldots, \phi_c(t)] = \pi \exp(t \ln(M(e)))$. Thus, $\delta_e$ is determined by $M(e)$ and $\pi$ (i.e. without knowing $R$ and $\lambda_e$ separately). If $\pi$ is the equilibrium vector, we recover (8). For two colors we can find $\delta_e$ explicitly:

$$\delta_e = \frac{1}{P} \left( P_{12}^2 \pi_1 + P_{21}^2 \pi_2 - P_{12} P_{21} - \frac{2P_{12} P_{21} \ln(1 - P)}{P} \right),$$

where $P = P_{12} + P_{21}$ and where $P_{ij} = p_s(i, j)$.

### 2.2.2. Group based models

If we regard $C$ as a group under some operation (written here multiplicatively), a group-based model places the following constraint on the transition matrices:

(G1) $p_e(x, \beta) = h_e(x^{-1} \beta)$ for all edges $e$ of $T^{+p}$,

where $h_e : C \rightarrow [0, 1]$ is defined for each edge $e$ of $T^{+p}$.

Thus, in a group-based model, we can think of a random group element $g(\rho)$ being assigned to $\rho$ according to the distribution $\pi = \pi(\rho)$ and a random group element $g(e)$ being assigned independently to each edge $e$ of $T^{+p}$ (according to the distributions $h_e$). Each leaf $i$ is then colored by the product (in the group) of $g(\rho)g(e_1) \cdots g(e_k)$ where $e_1, \ldots, e_k$ is the directed path from $\rho$ to $i$. It can be checked that the set of all transition matrices satisfying (G1) for a particular group structure on $C$ forms a semigroup, so that the set of group-based models are indeed a subset of the semigroup models.

When $C$ is a finite abelian group, Székely et al. [50] used discrete Fourier analysis to describe a relationship which gives the pattern probabilities in terms of the underlying tree with its associated functions $h_e$, and which is invertible under various restrictions on the $h_e$'s. Of particular interest are the elementary abelian 2-groups, $C_2^k$. These correspond to the 2-color Neyman model [10, 17, 39] (described at the beginning of Section 2.2) when $k = 1$, and to an extended version of Kimura’s 3ST model [32] (described in Section 2.2.3 below) allowing different rate matrices on different edges, when $k = 2$. Note that for a $C_2^k$-based model, the associated transition matrices form the multiplication table of the group $C_2^k$, since $p_e(x, \beta) = h_e(x^{-1} \beta) = h_e(x \beta)$.

**Definition.** Consider the following equivalence relation on patterns on $S$: $\chi_1$ is equivalent to $\chi_2$, if, for some element $g \in C$, and all leaves $i \in S$,

$$\chi_1(i) = \chi_2(i) g,$$

where multiplication is carried out in the group. Each equivalence class of patterns is called a quotient pattern, and the probability of a quotient pattern $\chi^\ast$ denoted $f_{\chi^\ast} = f_{\chi^\ast}(T^{+p}, h, \pi(\rho))$ is the probability of generating any pattern in the class (the sum of $f_\chi$ over all patterns $\chi$ in $\chi^\ast$), where $h := \{h_e\}$. Let $r$ be an arbitrary leaf of $T$. Note that the quotient patterns on $S$ are in 1–1 correspondence with the patterns on
$S' = S - \{r\}$; simply choose a color $\alpha$, and map each pattern $\chi$ on $S'$ to the equivalence class of the pattern $\chi_{x}$, where $\chi_{x}|S' = \chi$ and $\chi_{x}(r) = \alpha$.

By taking the quotient patterns for the $C_{2}^{k}$-based models we factor out the influence of the distribution $\pi(\rho)$ of the root $\rho$ and of its location on $T$. This result, which is formalized in the following theorem, is central to the next section, by allowing all calculations to be carried out on the unrooted tree $T$, rather than $T^{\rho}$. For each edge $e$ of $T$, let $h_{e}': C \to [0, 1]$ agree with $h_{e}$ if $e$ appears in $T^{\rho}$, otherwise if there is an edge $e'$ of $T$ that is bisected to form two edges $e_{1}, e_{2}$ in $T^{\rho}$ (i.e. in case $\rho$ has degree 2) let $h_{e}'$ be the convolution of $h_{e_{1}}$ and $h_{e_{2}}$, that is

$$h_{e}'(g) = \sum_{x \in C} h_{e_{1}}(x)h_{e_{2}}(x^{-1}g).$$

Let $f_{\rho}^{*}(T, h', \pi')$ be the probability of generating the quotient pattern $\chi^{*}$ under the model described above, when the root of the tree $T$ is taken to be leaf $r$, with associated color distribution $\pi'$, and where $h' := \{h_{e}'\}$.

**Theorem 5.** Let $\pi'$ be any distribution of colors at leaf $r$. Then,

$$f_{\rho}^{*}(T, h', \pi') = f_{\rho}^{*}(T^{\rho}, h, \pi(\rho)).$$

(Thus $f^{*}$ is independent of the distribution $\pi(\rho)$ of the colors at the root $\rho$, and of its location in $T$.)

**Proof.** Let $g(\rho)$ (resp. $g'(r)$) denote the random element of $C = C_{2}^{k}$ assigned to $\rho$ (resp. $r$) under $\pi(\rho)$ (resp. $\pi'$). Let $g(e)$ (resp. $g'(e)$) denote the random element of $C$ assigned to edge $e$ according to the distribution $h_{e}$ (resp. $h_{e}'$), and let $\chi(i)$ (resp. $\chi'(i)$) be the induced random color of leaf $i$, for $i \in S'$ by $(T^{\rho}, h, \pi(\rho))$ (resp. $(T, h', \pi')$). We have,

$$\chi(i) = g(\rho) \times \prod_{e \in P(T^{\rho}; \rho, i)} g(e).$$

For $i \in S'$, let $Z_{i} := \chi(i)\chi(r)$. Then,

$$Z_{i} = g(\rho) \times \prod_{e \in P(T^{\rho}; \rho, i)} g(e) \times \left( \prod_{e \in P(T^{\rho}; \rho, r)} g(e) \right) = \prod_{e \in P(T^{\rho}; i, r)} g(e), \quad (10)$$

since $C_{2}^{k}$ is abelian and $x^{2} = 1$ in this group. Similarly, for $i \in S'$, if we let $Z_{i}' := \chi'(i)$ $\chi'(r) = \chi'(i)g'(r)$ we have

$$Z_{i}' = \prod_{e \in P(T, i, r)} g'(e). \quad (11)$$

It follows from Eqs. (10) and (11) and the definition of $h_{e}'$ that the random vectors $Z := [Z_{i}; i \in S']$ and $Z' := [Z_{i}'; i \in S']$ have the same distribution. Now, the probability of any quotient pattern $\chi^{*}$ under $(T^{\rho}, h, \pi(\rho))$ (resp. under $(T, h', \pi')$) is simply the
probability that $Z_i$ (resp. $Z'_i$) equals $\chi_0(i)\chi_0(r)$ for all $i \in S'$, where $\chi_0$ is any pattern on $S$ in the equivalence class $\chi^*$. Since $Z$ and $Z'$ have the same distribution, we obtain the desired equality.

2.2.3. Inverting the extended Kimura 3ST model

Kimura [32] introduced a model he called the “three substitution-type (3ST) model”, where he assigned three rate parameters for nucleotide substitutions, $\alpha$ for the rate of transitions (A $\leftrightarrow$ G and T(U) $\leftrightarrow$ C), and $\beta$ and $\gamma$ for the two types of transversions. (The type I transversions (of rate $\beta$) are A $\leftrightarrow$ T(U) and G $\leftrightarrow$ C, and the type II (of rate $\gamma$) are A $\leftrightarrow$ C and T(U) $\leftrightarrow$ G.) He denotes $P, Q$ and $R$ as the probabilities of each of these substitutions between homologous sites of two sequences, descended from a common ancestral sequence over a time interval $t$. This set of nucleotide substitution types forms the Klein four group $C_2 \times C_2$, writing the element $(1,-1)$ to correspond to a transition, $(-1,1)$ and $(-1,-1)$ to correspond to transversions of type I and II respectively, together with $(1,1)$ representing no substitution [15].

Under Kimura’s stationary model the expected number of substitutions is

$$K = -\frac{1}{4} \ln\{(1 - 2P - 2Q)(1 - 2P - 2R)(1 - 2Q - 2R)\},$$

(12)

which is the sum of three components, being the expected numbers of each of the three substitution types,

$$Q_{+\cdot} = 2\alpha t = -\frac{1}{4} \ln\{(1 - 2P - 2Q)(1 - 2P - 2R)/(1 - 2Q - 2R)\},$$

$$Q_{-\cdot} = 2\beta t = -\frac{1}{4} \ln\{(1 - 2P - 2Q)(1 - 2Q - 2R)/(1 - 2P - 2R)\},$$

$$Q_{\cdot-} = 2\gamma t = -\frac{1}{4} \ln\{(1 - 2P - 2Q)(1 - 2Q - 2R)/(1 - 2P - 2Q)\},$$

writing $Q_{\cdot\cdot}$ for the expected number of changes corresponding to $(1,-1)$, etc.

We show below (Eq. (20)) that Eqs. (13) are easily inverted, so that the probabilities can be expressed as functions of the $Q_{ij}$ components. These $Q_{ij}$ components are linear with time for fixed rates under the stationary model, so they represent additive parameters for successive edges of a tree $T^{+\cdot}$. From Theorem 5 we recall that the nucleotide differences at the leaves of $T^{+\cdot}$ are independent of the root location and its color distribution, so we can apply our analysis to the associated unrooted tree $T$. We will refer to the $Q_{ij}$ components as “edge lengths” for each of the edges of $T$. In [44, 29] we considered the following generalization of Kimura’s 3ST model.

For each edge $e$ of $T$ (with $n$ leaves), we can specify three probabilities $P_{+\cdot}(e)$, $P_{\cdot\cdot}(e)$ and $P_{\cdot-}(e)$, (analogous to Kimura’s $P, Q$ and $R$), for the substitutions of each type across $e$. These will be determined from three edge length parameters $Q_{+\cdot}(e), Q_{\cdot\cdot}(e)$ and $Q_{\cdot-}(e)$. (Under a stationary model these $Q(e)$ values will be the expected numbers of changes of each type along edge $e$, but the relationship between the $P(e)$ and $Q(e)$ values does not depend on the assumption of stationarity).

Our spectral analysis [29] relates the set of quotient patterns at the leaves of the tree with the three edge lengths $Q_{ij}(e)$ for each edge of the tree, using elements of the
Hadamard matrix $H_{2n-2}$. ($H_k$ is a square matrix of $2^k$ rows, whose entries are $+1$ or $-1$, obtained by taking the $k$-fold Kronecker product of

$$H_1 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix},$$

so

$$H_k = \begin{bmatrix} H_{k-1} & H_{k-1} \\ H_{k-1} & -H_{k-1} \end{bmatrix},$$

and $H_k^{-1} = 2^{-k}H_k$.)

Once the edge lengths are known for all the edges of $T$, we are able to determine the probabilities of the quotient patterns at the leaves of $T$, which by Theorem 5, is independent of the root distribution. For each pair of subsets $\alpha, \beta$ of $S' = S - \{r\}$ let $s_{\alpha\beta}$ be the probability of the occurrence of the quotient pattern with

$$\alpha = \{i \in S' \mid (\chi(r)^{-1}\chi(i))_1 = -1\}, \quad \beta = \{i \in S' \mid (\chi(r)^{-1}\chi(i))_2 = -1\}.$$

(The suffixes refer to the first and second components of the elements of $V = C_2 \times C_2$. Thus $s_{\emptyset\emptyset}$ is the probability that all the leaves have the same color.)

We will refer to the vector (for a suitable ordering) $s$ of probabilities as the sequence spectrum. These $4^{n-1}$ probabilities are functions of the $3 \times k$ edge length parameters, where $k \leq 2n - 3$ is the number of edges of $T$. However, for convenience, we embed these $3k$ values in another vector of $4^{n-1}$ entries, called the edge length spectrum $q$, also indexed by pairs of subsets of $S'$. Let $E(T)$ be the set of edges of $T$. For the edge $e = e_x \in E(T), (\alpha \subseteq S')$ which induces the split $\alpha, S - \alpha$ on the leaves of $T$, let

$$q_{\emptyset\emptyset} = Q_{++}(e), \quad q_{\emptyset\emptyset} = Q_{-+}(e), \quad q_{\emptyset\emptyset} = Q_{--}(e).$$

We do this for each edge $e \in E(T)$. Set

$$q_{\emptyset\emptyset} = -\sum_{e \in E(T)} (Q_{++}(e) + Q_{-+}(e) + Q_{--}(e)),$$

and set all the remaining $q_{\alpha\beta}$ entries to 0. (Thus $\sum_{\alpha,\beta \subseteq S'} q_{\alpha\beta} = 0$ and the positive entries of the edge length spectrum define $T$.)

**Theorem 6** (Steel et al. [44])

$$s = H_{2n-2}^{-1} \exp(H_{2n-2}q), \quad (14)$$

where the exponential function is applied to each component of the vector individually.

The proof of this theorem is based on a more general result in [44, 49, 50] using group theory. Below we will give an outline of a proof which interprets some useful intermediate terms. However we will first introduce some useful corollaries.
Inverting Eq. (14) we obtain the edge length spectrum as a function of the sequence spectrum.

**Theorem 7**

\[ q = H_{2n-2}^{-1} \ln(H_{2n-2}s). \]  

(15)

where the logarithmic function is applied to each component of the vector individually.

If we index the entries of the Hadamard matrix \( H_{n-1} \) by the subsets of \( S' \), we find the entry

\[ h_{2\beta} = (-1)^{|\alpha \cap \beta|}, \quad \forall \alpha, \beta \subseteq S' \]  

(16)

and then Eqs. (14) and (15) can be usefully expressed [29] as

\[ s_{2\beta} = \frac{1}{4n-1} \sum_{x', \beta' \subseteq S'} h_{x'2h_{\beta'}} \exp \left( \sum_{x', \beta' \subseteq S'} h_{x'2h_{\beta'}} q_{x'2\beta'} \right) \quad \forall \alpha, \beta \subseteq S', \]  

(17)

and

\[ q_{2\beta} = \frac{1}{4n-1} \sum_{x', \beta' \subseteq S'} h_{2x'2h_{\beta'}} \ln \left( \sum_{x', \beta' \subseteq S'} h_{x'2h_{\beta'}} q_{x'2\beta'} s_{x'2\beta'} \right) \quad \forall \alpha, \beta \subseteq S'. \]  

(18)

In particular, with \( n = 2 \), Theorem 6 gives us the relationship between the \( s \) and \( q \) vectors for a single edge \( e \). Let \( P_{++}(e) \) be the probability that the endpoints of \( e \) have the same coloration (so \( \sum_{i,j \in \{+1,-1\}} P_{ij} = 1 \)), and let \( Q_{++}(e) = -K \) (so \( \sum_{i,j \in \{+1,-1\}} Q_{ij} = 0 \)), then with

\[ P(e) = \begin{bmatrix} P_{++} \\ P_{+} \\ P_{-} \\ P_{--} \end{bmatrix} = s \text{ and } Q(e) = \begin{bmatrix} Q_{++} \\ Q_{+} \\ Q_{-} \\ Q_{--} \end{bmatrix} = q, \]

\[ H_2 P(e) = \exp(H_2 Q(e)). \]

(19)

(It is useful also to note that the entries in \( H_2 P(e) \) are the eigenvalues of the transition matrix of nucleotide substitution across \( e \)).

**Proof of Theorem 6.** For any set \( X \) of edges of \( T \), let \( \chi(X) = \Pi_{e=(u,v) \in X} \chi(u)^{-1} \chi(v). \) (As \( \chi(u)^{-1} = \chi(u) \), the orientation of \( e \) is irrelevant.) For \( i, j \in \{+1, -1\} \) let \( P_{ij}(X) \) be the probability that \( \chi(X) = (i, j) \), and let \( P(X) \) be the vector of \( P_{ij}(X) \) values. Then, by Eq. (19) (which may easily be proved directly without recourse to Theorem 6) we have

\[ P(X) = H_2^{-1} \exp(H_2 Q(X)), \]

(20)
where \( Q(X) = \sum_{e \in X} Q(e) \). For any subset \( \alpha \subseteq S' \) we define a path set \( \Pi_\alpha \) as the disjoint union (symmetric difference) of the sets of edges in the paths from vertices \( i \in \alpha \) to \( r \) in \( T \). The set \( t_\alpha \) of leaf labels of \( \Pi_\alpha \) is an even ordered subset of \( S \), in particular

\[
t_\alpha = \begin{cases} 
\alpha & \text{when } |\alpha| \text{ is even,} \\
\alpha \cup \{ r \} & \text{is odd.}
\end{cases}
\]

We can determine the totality of substitutions across the edges of \( \Pi_\alpha \) by examining the product of the colors at the leaves. \( P_{ij}(\Pi_\alpha) \) is the probability that \( \prod_{u \in t_\alpha} z(u) = (i, j) \), which can be readily computed using Eq. (20).

We find by induction [26] that an edge \( e_{\alpha'} \) belongs to path set \( \Pi_{\alpha} \iff h_{e_{\alpha'}} = -1 \). Thus

\[
Q(\Pi_\alpha) = \sum_{h_{e_{\alpha'}} = -1} \begin{bmatrix} -(q_{\alpha'}^\beta + q_{\beta e_{\alpha'}} + q_{e_{\alpha'}^{\alpha'}}) \\
q_{\alpha'}^\beta \\
q_{\beta e_{\alpha'}} \\
q_{\alpha'}^{\alpha'} \end{bmatrix}.
\]

Consider the general term of \( \rho = Hq \):

\[
\rho_{\alpha\beta} = (Hq)_{\alpha\beta} = \sum_{\alpha', \beta' \subseteq S'} h_{\alpha\beta'}q_{\alpha'}^{\beta'}.
\]

All the terms of this sum are zero, except for \( q_{\emptyset \emptyset} \) and the three edge lengths for each edge in \( E(T) \). Thus

\[
\rho_{\alpha\beta} = q_{\emptyset \emptyset} + \sum_{e_{\alpha'} \in E(T)} (h_{\alpha\alpha'}q_{\emptyset \alpha'} + h_{\beta e_{\alpha'}}q_{\emptyset \alpha'} + h_{e_{\alpha'}^{\alpha'}}q_{\alpha'}^{\alpha'}),
\]

where \( \gamma = \alpha \cup \beta \) is the disjoint union of \( \alpha \) and \( \beta \). However, as \( -q_{\emptyset \emptyset} \) is the sum of all the other \( q \) terms, we find

\[
\rho_{\alpha\beta} = -2[(Q_{-\alpha})(\Pi_\alpha) + Q_{+\alpha}(\Pi_\beta) + Q_{-\alpha}(\Pi_\gamma)].
\]

We now rearrange the terms of Eq. (22), noting that \( \Pi_\alpha \) can be partitioned into \( X \cup Z \), \( \Pi_\beta \) can be partitioned into \( Y \cup Z \), and \( \Pi_\gamma \) can be partitioned into \( X \cup Y \), where \( X = \Pi_\alpha - \Pi_\beta, Y = \Pi_\beta - \Pi_\alpha \), and \( Z = \Pi_\alpha \cap \Pi_\beta \). Hence Eq. (22) can be expressed as

\[
\rho_{\alpha\beta} = -2[(Q_{-\alpha}(X) + Q_{-\alpha}(X))] + [Q_{-\alpha}(Y) + Q_{+-\alpha}(Y)] + [Q_{-\alpha}(Z) + Q_{+\alpha}(Z)]).
\]

Taking the exponential we obtain the product of three terms. The first term is

\[
\exp(-2[Q_{-\alpha}(X) + Q_{-\alpha}(X)]) = [P_{++}(X) - P_{+-}(X) + P_{-+}(X) - P_{--}(X)]
\]

by Eq. (15), which we can write as

\[
\exp(-2[Q_{+\alpha}(X) + Q_{-\alpha}(X)]) = \sum_{a,b \in \{-, +\}} aP_{ab}(X).
\]
Likewise we can express the other two terms as

$$\exp(-2[Q_{+-}(Y) + Q_{-+}(Y)]) = \sum_{c,d \in \{-, +\}} dP_{cd}(Y),$$

$$\exp(-2[Q_{+-}(Z) + Q_{-+}(Z)]) = \sum_{e,f \in \{-, +\}} efP_{ef}(Z).$$

The three sets $X, Y$ and $Z$ are each disjoint, so the probabilities can be multiplied together to give

$$\left(\exp Hq\right)_{\beta\beta} = \left(\sum_{a, b \in \{-, +\}} aP_{ab}(X)\right) \times \left(\sum_{c, d \in \{-, +\}} dP_{cd}(Y)\right) \times \left(\sum_{e, f \in \{-, +\}} efP_{ef}(Z)\right),$$

$$= \sum_{a, b, c, d, e, f \in \{-, +\}} adefP_{ab}(X)P_{cd}(Y)P_{ef}(Z),$$

$$= \sum_{a, b, c, d, e, f \in \{-, +\}} adefP_{as}(X)P_{sd}(Y)P_{ef}(Z),$$

where we write $P_{as}(X) = P_{a-}(X) + P_{a+}(X)$, etc. The terms of $s$ which contribute to $P_{as}(X)P_{sd}(Y)P_{ef}(Z) = P_{as}(\Pi_x - \Pi_{\beta})P_{sd}(\Pi_{\beta} - \Pi_x)P_{ef}(\Pi_x \cap \Pi_{\beta})$ are the terms $s_{x', \beta'}$ where

$$h_{x'(\beta-\beta)} = a, \quad h_{\beta'(\beta-\beta)} = d, \quad h_{x'(\beta \cap \beta)} = e, \quad \text{and} \quad h_{\beta'(\beta \cap \beta)} = f. \quad (24)$$

Hence, noting that $h_{x'(\beta-\beta)}h_{x'(\beta \cap \beta)} = h_{x'\beta}$ and $h_{\beta'(\beta-\beta)}h_{x'(\beta \cap \beta)} = h_{\beta'\beta}$:

$$\left(\exp Hq\right)_{\beta\beta} = \sum_{a, d, e, f \in \{-, +\}} \sum_{x', \beta' \in \{1, \ldots, n\}} h_{x'\beta}h_{\beta'\beta'}s_{x'\beta'},$$

where the inner sum $\sum_{x', \beta' \in \{1, \ldots, n\}}$ is over all pairs of subsets $x', \beta'$ which satisfy Eq. (24) for the given values of $x, \beta, a, d, e, f, \beta'$. These conditions ensure that all 16 sign combinations from Eq. (24) are met, so changing the order of summation, we sum over all pairs of subsets to obtain

$$\left(\exp Hq\right)_{\beta\beta} = \sum_{x', \beta' \subseteq S'} h_{x'\beta'}h_{\beta'\beta'}s_{x'\beta'}, \quad (25)$$

from which the theorem follows. We had initially applied spectral analysis to the two-color Neyman model [26], which is described at the start of Section 2.2. This case can be obtained from Theorem 6 by setting the probability of all transversions to 0. Then
we are able to express the results in a simpler manner, where we write $s_x$ for $s_{x\emptyset}$ and $q_x$ for $q_{x\emptyset}$ so both become vectors of $2^{n-1}$ components indexed by the subsets of $S'$. □

**Theorem 8** (Hendy [26])

$$s = H_n^{-1} \exp(H_{n-1}q),$$

and

$$q = H_n^{-1} \ln(H_{n-1}s).$$

These equations can also be interpreted in terms of pathsets where for $\beta \subseteq S'$ let

$$\rho_\beta = \sum_{x \subseteq S'} h_{x\beta}q_x = -2 \sum_{e_i \in \Pi_\beta} q_{e_i},$$

which we define as the length of $\Pi_\beta$. (Under a stationary model $\rho_\beta$ is the expected number of changes along the edges of pathset $\Pi_\beta$.) Thus from Theorem 8,

$$s_x = \sum_{\beta \subseteq S'} h_{\beta x} \exp(\rho_\beta).$$

### 2.3. An extension: variable rates across sequence sites

Suppose that the substitution process is according to the generalized Kimura 3ST model, but is proceeding at different rates across the sites. For convenience, we let $C(T) = \bigcup_{e \in \sigma(T)} \{\emptyset, e, \emptyset, \Sigma\}$. Recall that for $\theta \in C(T)$ (with associated $\gamma \in \sigma(T)$), $q_\theta$ is the expected number of one of the three types of substitution (considered by Kimura) on the edge(s) of $T$ which induce the split $(\gamma, S - \gamma)$ in $T$. Suppose that this quantity varies from site to site in the sequences: then let $q_\theta(j)$ denote the expected value of $q_\theta$ at site $j$. We consider the more general model specified by the condition:

$$q_\theta(j) \text{ can be written in the form } q_\theta \times \lambda_j.$$

Here $\lambda_j$ can be thought of as the rate at which substitutions occur at site $j$, and $q_\theta$ is the average (over all the sites) of the expected number of the particular type of substitution on the edge of $T$ corresponding to $\theta$, divided by the average value of the $\lambda_j$'s. This type of “geometric” scaling model is also considered by Chang [11].

Let $\mu(x) = (1/m) \sum_{j=1}^m \exp(x\lambda_j)$, the average value of the numbers $\exp(x\lambda_j)$ (averaged over all sites $j$). Thus, if the rate parameters $\lambda_j$ are drawn independently according to some distribution, then $\mu(x)$ is approximated by the moment generating function of this distribution. Now, $\lambda_j$ is positive for all $j$, and so $\mu(x)$ is monotone increasing, and therefore has a unique left functional inverse, $\mu^{-1}(x)$, so that $\mu^{-1}(\mu(x)) = x$, $\forall x \in \mathbb{R}$. Let $\bar{s}$ be the average value of the sequence spectrum across the sites. Then we have the following result, where $\mu$ and $\mu^{-1}$ are applied componentwise on vectors.
Theorem 9 (Steel et al. [46]). For the extended 3ST model with underlying tree $T$ and arbitrary root distribution, we have

$$\tilde{s} = H_{2n-2}^{-1} \mu(H_{2n-2} 2q),$$

(28)

and

$$q = H_{2n-2}^{-1} \mu^{-1}(H_{2n-2} \tilde{s}).$$

(29)

Example. (1) In the case that all the sites evolve at the same rate ($= \lambda$), we have $\mu(x) = \exp(\lambda x)$, giving $\mu^{-1}(x) = (1/\lambda) \ln(x)$, and so Theorem 7 is just a special case of Theorem 9.

(2) Jin and Nei [31] suggest that the gamma distribution

$$f(x) = \frac{\exp(-vx)x^{k-1}v^k}{\Gamma(k)}, \quad x > 0,$$

may approximate the distribution of the $\lambda_j$. In this case,

$$\mu(x) = \left( \frac{v}{v - x} \right)^k$$

so that $\mu^{-1}(x) = v(1 - \phi_k(x))$, where $\phi_k(x) = x^{-1/k}$, and so

$$(e_1 - (H_{2n-2}^{-1} \phi_k(H_{2n-2} \tilde{s}))_{\alpha \beta} = 0 \iff \alpha \beta \in C(T) \cup \{0\},$$

where $e_1 = [1, 0, 0, 0, \ldots]^T$.

Remark. (1) Theorem 9 shows that, provided the distribution of rates across sites is known, then $T$ can be recovered from $\tilde{s}$. Indeed this holds under certain conditions, even if the distribution (or $\tilde{s}$) is not known exactly, but suitably constrained [47]. However if the distribution is variable, it is possible for all trees to give identical $\tilde{s}$ values (and hence render consistent tree reconstruction impossible by any method) by suitable choices of distributions and edge parameters for each tree (for details, see [47]).

(2) In the case of a stationary model, a distribution of rates across sites is modelled by taking the rate matrix for the process at site $j$ to be $R \times \lambda_j$. In that case, provided $R$ forms a reversible model, the transformation

$$\delta_{ij} = - \text{tr}(H \mu^{-1}[\Pi^{-1}F_{ij}])$$

(where, $\mu^{-1}$ is as above, but applied to matrices, and $F_{ij}$ is the divergence matrix described in Section 2.1) is additive on the true tree, and recovers the underlying unrooted tree (for details see [52] or [23]).
3. Applications

Theorems 6–8 describe correspondences between the edge length spectrum and the sequence spectrum under the two or four color models described. From the edge length spectrum, the tree $T$ can be identified. There are a number of applications that can be made from these relationships, which we now list.

Firstly, for the analysis of sequence data, we use an observed set of frequencies $\hat{s}$ of patterns from sequence data as an estimate of the probabilities $s$, then provided the logarithms exist, Theorems 7 and 8 can be used to calculate an estimate $\gamma = H^{-1}\ln(H\hat{s})$ of the edge length spectrum. We refer to this transformation of the observed data as a Hadamard conjugation and to $\gamma$ as the conjugate spectrum. Various fitting procedures can be used to estimate $T$ from $\gamma$. In [26] we introduce a least-squares procedure called the closest tree procedure. This procedure estimates the edge length spectrum, $q$, from which $s = H^{-1}\exp(H\gamma)$ can be calculated and compared to the observed frequencies $\hat{s}$; see [29] for an application to biological data. Alternatively, as the entries of $\gamma$ represent “corrected” lengths, a traditional method, for example maximum parsimony, can be applied [45].

Another application of Theorems 6–8 is the derivation and classification of all the “phylogenetic invariants” (polynomial functions of the pattern probabilities which for some associated phylogenetic tree take the value 0 for any choices of the matrices $M(e)$). Furthermore, one can classify all linear invariants for various submodels of the Kimura 3ST model, and the 2-color Neyman model (both with and without a molecular clock). In particular, the dimensions of the vector space of linear invariants for these models can be conveniently detailed by formulae that in some cases just involve the number of leaves ($n$) and Fibonacci numbers. (For details on linear invariants see [20, 28], and for the classification of non-linear invariants for the Kimura 3ST model, see [46] and [15]).

A third application of Theorems 6 and 8 is to analyze various phylogenetic tree building methods. We can generate sample sequence frequencies from a known tree $T$ and specified edge lengths. Samples can then be used to test the accuracy of the method. Sometimes methods are inconsistent under a particular model of sequence generation, i.e., the methods do not improve with accuracy as the sampling error is reduced by using longer sequences, leading to the situation that the incorrect tree is always found when the sampling error is zero. Felsenstein [18] showed that the popular maximum parsimony method (applied to $\hat{s}$) can be inconsistent, even under the 2-color Neyman model with only four taxa. In his example the molecular clock hypothesis was violated. In [27] the Hadamard conjugation was used to show that for the 2-color Neyman model under the molecular clock hypothesis, maximum parsimony must be consistent for four taxa, but it can be inconsistent with five or more taxa. As an illustration of the usefulness of Theorem 6 we show below that maximum parsimony on $\hat{s}$ is also consistent with four taxa under Kimura’s 3ST model and the molecular clock.
3.1. Consistency of maximum parsimony on four colors and four taxa with the molecular clock hypothesis

Let $T^{+\rho}$ be a fully resolved phylogenetic tree on leaf set $S = \{1, 2, 3, 4\}$. We may assume that the associated unrooted phylogenetic tree $T$ is the tree $T_1$ shown in Fig. 2; the other two unrooted fully resolved phylogenetic trees on leaf set $S$ are also shown as $T_2$ and $T_3$.

Suppose we have a phylogenetic tree $T$ on leaf set $S$, together with a collection of aligned sequences, one for each species in a set $S$, and thereby inducing a collection of patterns on $S$. For each site the Fitch length [19] is the minimum number of edges of $T$ which must be assigned differently colored ends in any extension of the leaf coloration (of the pattern induced by the site) to all vertices of $T$. The parsimony length for $T$ is the sum of these Fitch lengths over all the sites. We define the length of $T$, $l(T)$, to be the average of the Fitch length over all sites. The maximum parsimony tree is the unrooted fully resolved tree $T$ with the smallest parsimony length and hence the tree for which $l(T)$ is minimal. Hence for a set of four taxa, $T_1$ is the maximum parsimony tree $\Leftrightarrow l(T_1) < l(T_2)$ and $l(T_1) < l(T_3)$.

Theorem 6 gives the expected frequency of sites with pattern $(x, \beta)$ as $s_{x\beta}$. For a given tree $T$, and i.i.d. sequence site evolution, the expected length $\bar{l}(T)$ is the expected value of $l(T)$ under the extended Kimura 3ST model, and so is the sum, over all patterns $(x, \beta)$, of $s_{x\beta}$ times the Fitch length for that pattern on $T$. Most patterns have the same Fitch length on each tree, and for trees on four taxa the differences between their expected lengths are a combination of the frequencies of only 6 of the 64 terms. Hence for example with $x = \{1, 2\}, \beta = \{1, 3\}$ and $\gamma = \{2, 3\},$

$$\bar{l}(T_2) - \bar{l}(T_1) = (s_{\emptyset x} + s_{x\emptyset} + s_{x\gamma}) - (s_{\emptyset \beta} + s_{\rho \emptyset} + s_{\rho \beta}). \tag{30}$$

Thus $T_1$ is the maximum parsimony tree $\Leftrightarrow$

$$(s_{\emptyset x} + s_{x\emptyset} + s_{x\gamma}) - (s_{\emptyset \beta} + s_{\rho \emptyset} + s_{\rho \beta}) > 0$$

and

$$(s_{\emptyset x} + s_{x\emptyset} + s_{x\gamma}) - (s_{\emptyset \gamma} + s_{\gamma \emptyset} + s_{\gamma \gamma}) > 0.$$

The molecular clock hypothesis states that a "time scale" can be applied to the edges of $T^{+\rho}$, so that the expected numbers of color changes on an edge are proportional to this time. Specifically for each vertex $v$, we assign a parameter "time" $t(v)$. (Biologically this refers to the historical time that the bifurcation event at $v$ occurred.) For each edge $e = (u, v)$ of $T^{+\rho}$ we define the time span $t_e = |t(u) - t(v)|$. (Thus the sum of the time spans on the path from $\rho$ to any leaf is constant.) For the edge $e_\rho = (w, x)$ of $T$ where $w$ and $x$ are adjacent to $\rho$ in $T^{+\rho}$, we define $t_\rho = |2t(\rho) - t(w) - t(x)|$. 

In Kimura’s 3ST model of nucleotide evolution this implies that there are three parameters, \( \lambda_1, \lambda_2, \lambda_3 > 0 \) so for each edge \( e_x \) of \( T \),

\[
q_{x0} = \lambda_1 t_x, \quad q_{0x} = \lambda_2 t_x, \quad q_{xx} = \lambda_3 t_x,
\]

to give the edge length spectrum for \( T \).

**Theorem 10.** Maximum parsimony is consistent under Kimura’s 3ST model for four taxa with the molecular clock hypothesis.

**Proof.** There are two cases to consider: (1) the root \( \rho \) is on the central edge \( e_{\{1,2\}} \) of \( T_1 \) or (2) the root \( \rho \) is on one of the pendant edges, \( e_{\{1,2,3\}} \), say.

In case (1), the molecular clock hypothesis implies that there are three independent time parameters \( t_1, t_2, \) and \( t_3 \), where \( t_1 = t_{\{1\}} = t_{\{2\}}, t_2 = t_{\{3\}} = t_{\{1,2,3\}} \), and \( t_3 = t_{\{1,2\}} \geq |t_1 - t_2| \). Applying Theorem 6 to (30) we can show

\[
\tilde{t}(T_2) - \tilde{t}(T_1) = \frac{1}{8} \sum_{a,b,c} \exp(-2(b + c)(t_1 + t_2)) \times [1 - \exp(-2(b + c)t_3)]\left[1 + \exp(-4a(t_1 + t_2))\right] + \frac{1}{16} \sum_{a,b,c} \exp(-2(b + c)t_1 - \exp(-2(b + c)t_2))^2 - \exp(-2(bt_1 + ct_2)) + \frac{1}{16} \sum_{a,b,c} \exp(-4a(t_1 + t_2))\left[\exp(-2(bt_1 + ct_1)) - \exp(-2(bt_2 + ct_1))\right]^2,
\]

where the sums are over the three even permutations \((a,b,c)\) of \((\lambda_1, \lambda_2, \lambda_3)\). As each of the exponentials lie in the interval \((0,1)\), each term in the sums is positive, and \( \tilde{t}(T_2) > \tilde{t}(T_1) \). As \( t_{\{1\}} = t_{\{2\}} \) we can interchange vertices 1 and 2 to find \( \tilde{t}(T_3) = \tilde{t}(T_2) \). Hence \( T_1 \) has the smallest expected length.

In case (2), with the root on edge \( e_{\{1,2,3\}} \), the molecular clock hypothesis implies that there are three independent time parameters \( t_1, t_2, \) and \( t_3 \), where \( t_1 = t_{\{1\}} = t_{\{2\}}, t_2 = t_{\{1,2,3\}}, \) and \( t_3 = t_{\{1,2\}} \) and \( t_{\{3\}} = t_1 + t_3 \leq t_2 \). Hence from (30)

\[
\tilde{t}(T_2) - \tilde{t}(T_1) = \frac{1}{16} \sum \exp(-4(b + c)t_1)[1 - \exp(-4(b + c)t_3)] + \frac{1}{16} \sum \exp(-4at_1 - 2(b + c)(t_1 + t_2 + t_3)) \times [1 - \exp(-4at_3)][\exp(-4bt_1) + \exp(-4ct_1)],
\]

where again the sums are over the three even permutations \((a,b,c)\) of \((\lambda_1, \lambda_2, \lambda_3)\). As in case (1), each of the exponentials lie in the interval \((0,1)\), so each term in the sums is positive, and \( \tilde{t}(T_2) > \tilde{t}(T_1) \). As \( t_{\{1\}} = t_{\{2\}} \) interchanging vertices 1 and 2 gives \( \tilde{t}(T_3) = \tilde{t}(T_2) \). Hence \( T_1 \) has the smallest expected length.
Thus, in both cases, as the sequence length tends to infinity, the probability that the maximum parsimony tree is the true tree tends to 1. This is the condition for statistical consistency. □

We now provide a second application of Hadamard conjugation to the analysis of the maximum parsimony method.

3.2. A Poisson-style bound for the histogram of Fitch lengths under the 2-color Neyman model

Suppose the underlying phylogenetic tree $T$ is fully resolved. For a site that evolves on $T$, the Fitch length of this site on $T$ will always be less than or equal to the true number of substitutions that occurred on $T$ in creating the pattern observed at the leaves. Thus if the substitution probability on all the edges of the tree is small we would expect the histogram of the numbers of sites versus their Fitch length to fall off rapidly as the Fitch length increases, since a site with a large Fitch length must have required a large number of (improbable) substitutions. Unfortunately with real data we do not have the privilege of viewing the substitution probabilities, but we would still like to make predictions regarding the histogram of Fitch lengths. Here we show that just the first two entries of this histogram (i.e. the expected number of sites of Fitch length 0 (constant sites) and Fitch length 1) place constraints on the rate of decay of the remainder of the histogram – regardless of the unknown parameters on the underlying tree, when the sites evolve under the 2-color Neyman model (described at the start of Section 2.2). For sequences in which some sites are invariant (cannot undergo substitution) while the remaining sites evolve i.i.d. according to the 2-color Neyman model, Theorem 11 predicts a lower bound on the number of invariant sites (see [35] for an application).

Let $P[k]$ denote the probability of generating under the 2-color Neyman model on tree $T^{+\rho}$, a pattern with Fitch length $k$ on $T$, and let $P^*[k] := \sum_{j \geq k} P[j]$, the probability of generating under the 2-color Neyman model on tree $T^{+\rho}$, a pattern having Fitch length at least $k$ on $T$. Let $\mu = P[1] / P[0]^2$.

**Theorem 11**

$$P^*[k] \leq P[1] \sum_{j \geq k} \frac{\mu^{j-1}}{j!}.$$

**Proof.** First note that since $T$ is fully resolved, and we may assume $\rho$ has degree 2, $T^{+\rho}$ must have precisely $2n - 2$ edges, where $n = |S|$. Let $R[k]$ denote the probability that there are exactly $k$ edges of $T^{+\rho}$ on which a color change occurs under the 2-color Neyman model. Let $E(T^{+\rho})$ be the set of edges of $T^{+\rho}$. Thus,

$$R[j] = \sum_{U: |U| = j} \left( \prod_{v \in U} p_v \prod_{e \in E(T^{+\rho}) \setminus U} (1 - p_e) \right), \quad (31)$$
where $p_{e}$ is the probability of a color change on edge $e$, and the first summation is over all subsets $U$ of edges of $T^{+\rho}$ of cardinality $j$. Let

$$v = \sum_{e \in E(T^{+\rho})} \frac{p_{e}}{1 - p_{e}},$$

then, by (31),

$$R[1] = R[0]v \text{ and } R[j] \leq R[0] \frac{v^{j}}{j!}.$$  \hspace{1cm} (32)

A pattern of Fitch length 1 on $T$ is always generated whenever exactly one edge of $T^{+\rho}$ has a color change (under the model), thus $P[1] \geq R[1]$; also, a pattern with Fitch length at least $k$ on $T$ requires at least $k$ edges to have color changes on $T^{+\rho}$ in its generation (under the model) so $P^{*}[k] \leq \sum_{j \geq k} R[j]$. Combining these two inequalities with the two inequalities in (32), we deduce that

$$P[1] \geq R[0]v,$$  \hspace{1cm} (33)

$$P^{*}[k] \leq R[0] \sum_{j \geq k} \frac{v^{j}}{j!} \leq P[1] \sum_{j \geq k} \frac{v^{j-1}}{j!}.$$  \hspace{1cm} (34)

We claim

$$P[0]^{2} \leq R[0].$$  \hspace{1cm} (35)

The theorem then follows; since combining (33) and (35), gives $v \leq P[1]/P[0]^{2}$, which together with (34) gives the theorem.

We now proceed to establish (35). For each edge $e$ of $T^{+\rho}$, let $q_{e} := -\frac{1}{2} \ln(1 - 2p_{e})$, then,

$$R[0] = \prod_{e \in E(T^{+\rho})} (1 - p_{e}) = \prod_{e \in E(T^{+\rho})} \frac{1 - \exp(-2q_{e})}{2} = \frac{1}{2^{2n-2}} \times \sum_{U \subseteq E(T^{+\rho}), e \in U} \prod_{e \in E(T^{+\rho})} \exp(-2q_{e}),$$

where the summation is over all subsets $U$ of the $2n - 2$ edges of $T^{+\rho}$. Thus,

$$R[0] = \frac{1}{2^{2n-2}} \sum_{U \subseteq E(T^{+\rho})} \exp \left( -2 \sum_{e \in U} q_{e} \right).$$  \hspace{1cm} (36)

We now invoke Theorem 8. First let $e_{1}$ and $e_{2}$ denote the two edges of $T^{+\rho}$ incident with $\rho$. For $z \in \sigma(T)$, let $q_{z} = q_{e}$, if $(z, S - z)$ is the split obtained from $T^{+\rho}$ by deleting an edge $e \neq e_{1}, e_{2}$, and set $q_{z} = q_{e_{1}} + q_{e_{2}}$ otherwise. Extending $q_{z}$ to all subsets $z$ of $S'$ following the terminology of Eq. (27), $P[0]$ is just $s_{0}$ for $T$ with this associated vector $q$. Thus, from Theorem 8,

$$P[0] = s_{0} = \frac{1}{2^{n-1}} \sum_{\beta \subseteq S'} \exp(\rho_{\beta}),$$
where \( \rho_\beta \) is given in Eq. (26). Thus,

\[
P[0]^2 = \frac{1}{2^{2n-2}} \sum_{\beta, \beta' \subseteq S'} \exp(\rho_\beta + \rho_{\beta'}),
\]

and so that, from Eq. (26)

\[
P[0]^2 \leq \frac{1}{2^{2n-2}} \sum_{\beta, \beta' \subseteq S'} \exp \left( -2 \sum_{x \in \Pi_\beta \cup \Pi_{\beta'}} q_x \right),
\]

where \( n = |S| \) and \( \Pi_\beta, \Pi_{\beta'} \) are pathsets as defined in the proof of Theorem 6.

Now, for any rooted fully resolved tree \( T^{+\rho} \) there is a bijection \( \Psi \) from \( 2^S \times 2^S \) to the set of subsets of \( E(T^{+\rho}) \) (the edge set of \( T^{+\rho} \)) such that \( \Psi(\beta, \beta') \) is a subset of \( \beta \cup \beta'. \forall \beta, \beta' \in S \) (such a bijection can be constructed recursively). This bijection shows that each summation term in (37) is less than a corresponding summation term in (36), hence \( P[0]^2 \leq R[0] \), as required, completing the proof. \( \square \)

4. Conclusion

Under very few restrictions on the stochastic process of character substitution on a phylogenetic tree \( T^{+\rho} \), the discrete structure of \( T \) can be recovered from the expected frequencies of colorings at its leaves. This is important for phylogenetic inference, as it shows that for sufficient data, the tree is potentially recoverable from observable sequence data. The assumption that sites evolve at identical rates can be weakened and the conclusion is still valid in some cases. However the invertible relationships are between probabilities; any real data will be from finite samples and hence the effects of sampling need to be considered. These issues have recently begun to be seriously addressed, see [14, 16, 51]. It is apparent that the edge lengths must not be too small, or too large, as in that case, errors induced by sampling can be dominant. There are of course other potential complications influencing the accuracy of such inference, such as the validity of the i.i.d. assumption and the possibility of data error.

The other main focus of this paper has been Hadamard conjugation, and its applications. It is likely that many more applications of this technique can be found, particularly for analyzing the performance of different phylogenetic methods under suitable models.

Acknowledgements

We thank Chris Tuffley for some helpful criticism of a draft of this manuscript, and the referees for their useful suggestions and comments.
References