LETTER TO THE EDITOR

$U(1) \times U(1) \times U(1)$ symmetry of the Kimura 3ST model and phylogenetic branching processes

J D Bashford[1], **P D Jarvis**[1], **J G Sumner**[1] **and M A Steel**[2]

[1] School of Mathematics and Physics, University of Tasmania, GPO Box 252-21,
Hobart Tas 7001, Australia
[2] Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand

Abstract
An analysis of the Kimura 3ST model of DNA sequence evolution is given on
the basis of its continuous Lie symmetries. The rate matrix commutes with a
$U(1) \times U(1) \times U(1)$ phase subgroup of the group $GL(4)$ of 4×4 invertible
complex matrices acting on a linear space spanned by the four nucleic acid base
letters. The diagonal 'branching operator' representing speciation is defined,
and shown to intertwine the $U(1) \times U(1) \times U(1)$ action. Using the intertwining
property, a general formula for the probability density on the leaves of a binary
tree under the Kimura model is derived, which is shown to be equivalent to
established phylogenetic spectral transform methods.

PACS numbers: 87.23.Kg, 89.75.Hc, 02.50.Ey, 02.50.Ga, 03.65.Fd

The use of Markov models of stochastic change to taxonomic character distributions is
part of the standard armoury of techniques for describing mutations and inferring ancestral
relationships between taxa. For the simplest models, symmetries of the rate matrix under
discrete group actions ($\mathbb{Z}_2$ for binary types, or $\mathbb{Z}_2 \times \mathbb{Z}_2$ for DNA or RNA bases in molecular
applications, for example) have been used to good effect in simplifying phylogenetic analysis
(for references, see below). In particular, much attention has been centred on properties of the
frequently used Kimura 3ST model [1] which possesses such symmetry.

A general framework for phylogenetic branching models is as follows [2]. By assumption,
different taxonomic units are identified, and classified by a set of defining characteristics: for
example, based on morphological features or on sequence data, say, for a particular gene
or protein. To each taxon is ascribed a character probability density, and it is the task of
phylogenetic reconstruction to infer ancestral relationships amongst a group of related taxa,
given sample character frequencies.

In this letter, we describe an approach to the analysis of symmetries of such models
using *continuous* transformation groups. Rather than identifying the character types with

elements of a (non-Abelian or Abelian) discrete 'colour' group which patterns the rate matrix for transitions between types into orbit classes (see [3] and references therein), we look at linear transformations on the 'character space' spanned by the character types, and consider (complex, invertible) matrices which *commute* with the rate matrix. As we shall show, this approach, when implemented in the Kimura model, leads to an analysis which is well adapted to the established Hadamard discrete Fourier transform formalism [3–6], but which importantly has potential generalizations going beyond the binary colour groups.

Formally, let $\{p_a(t), a = 1, \ldots, K\}$ be the theoretical probabilities that the system has character $a = 1, 2, \ldots, K$, respectively. Introducing unit vectors $e_a, a = 1, \ldots, K$ the state vector representing the system[3]

$$p(t) = p_1(t)e_1 + p_2(t)e_2 + \cdots + p_K(t)e_K \tag{1}$$

is subject to linear time evolution[4]

$$\frac{\mathrm{d}}{\mathrm{d}t} p(t) = \widehat{R} \cdot p(t) \tag{2}$$

where the operator $\widehat{R}$ is a suitable $K \times K$ Markov rate matrix. It is natural to decompose $\widehat{R}$ as

$$\widehat{R} = \lambda(-\mathbb{1} + \widehat{T}) \tag{3}$$

where the traceless part $\widehat{T}$ belongs by definition to the Lie algebra $sl(K)$ (see below for the $K = 4$ case). The usual (positive) rates for substitution between different characters are thus the off-diagonal elements of $\widehat{T}$. A formal solution to (2) for time-independent rates is

$$p(t) = \mathrm{e}^{-\lambda t} \cdot \mathrm{e}^{\lambda t \widehat{T}} \cdot p(0). \tag{4}$$

The vector $p(t)$ (*a priori* in $\mathbb{C}^K$) is a probability density if each p_a is real, $p_a \geqslant 0$ and $\sum_a p_a = 1$. Consistent with the time dependence imposed by the master equation, given a starting density, probability conservation is implemented by demanding that $\widehat{R}$ is a unit column sum matrix. Introducing the vector Ω representing the sum of all the unit vectors in the distinguished basis

$$\Omega = e_1 + e_2 + \cdots + e_K$$

probability conservation requires that the dual $\Omega^\perp$ (the row vector with unit entries) is annihilated by $\widehat{R}$ regarded as an operator on the dual space, $\Omega^\perp \cdot \widehat{R} = 0$. Equivalently, $\Omega^\perp$ is a left *unit* eigenvector of $\widehat{T}$.

In the Kimura 3ST model [1] the characters a are of course the standard nucleic acid base letters A, G, U and C, and the rate matrix is[5]

$$\begin{bmatrix} \widehat{R}_{AA} & \widehat{R}_{AG} & \widehat{R}_{AU} & \widehat{R}_{AC} \\ \widehat{R}_{GA} & \widehat{R}_{GG} & \widehat{R}_{GU} & \widehat{R}_{GC} \\ \widehat{R}_{UA} & \widehat{R}_{UG} & \widehat{R}_{UU} & \widehat{R}_{UC} \\ \widehat{R}_{CA} & \widehat{R}_{CG} & \widehat{R}_{CU} & \widehat{R}_{CC} \end{bmatrix} = -(\alpha + \beta + \gamma)\mathbb{1} + \begin{bmatrix} 0 & \alpha & \beta & \gamma \\ \alpha & 0 & \gamma & \beta \\ \beta & \gamma & 0 & \alpha \\ \gamma & \beta & \alpha & 0 \end{bmatrix} \tag{5}$$

[3] Since the description is for a species or population, the possible interdependence of characters (as in the heritability of traits amongst individuals) is not directly addressed at this level but becomes an empirical issue of the appropriateness of the choice of characters.

[4] In order to make the models tractable, the rate matrix is assumed constant for each taxon, although the parameters can be adjusted between taxa (to allow for different metabolic rates for example), as described below. The time-dependent case is treated in [7] in a different context.

[5] The rate parameters α, β, γ describe base transitions, and two classes of transversions, respectively. The 3ST model is but one member of a hierarchy of base substitution models [8], which includes non-symmetric cases which attempt to account for properties such as differing base pair binding energies. Despite its simplicity, the 3ST model does describe some datasets adequately (see, for example [5], and the concluding remarks below).

wherein the total change rate parameter in (3) above is $\lambda = \alpha + \beta + \gamma$, and the traceless part of the rate operator can be written in the form

$$\widehat{T} = \frac{\alpha}{\alpha + \beta + \gamma}\widehat{K}_\alpha + \frac{\beta}{\alpha + \beta + \gamma}\widehat{K}_\beta + \frac{\gamma}{\alpha + \beta + \gamma}\widehat{K}_\gamma. \tag{6}$$

Remarkably the three Kimura matrices

$$\widehat{K}_\alpha = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix} \qquad \widehat{K}_\beta = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \qquad \widehat{K}_\gamma = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix} \tag{7}$$

provide a *maximal set of commuting generators* for the Lie algebra $sl(4)$ and thus can be chosen as the basis for a Cartan subalgebra; equivalently there exists a transformation of the basis spanned by e_A, e_G, e_U and e_C onto a new basis, in which the Kimura generators are diagonal (with doubly degenerate eigenvalues ± 1 by the traceless property, and the fact that they are square roots of $\mathbb{1}$). This transformation is well known to be generated by the Hadamard matrix H, which sends $\widehat{K}_i$ as matrices to $\mathsf{H}\widehat{K}_i\mathsf{H}^{-1}$, $i \in \{\alpha, \beta, \gamma\}$:

$$\mathsf{H} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \qquad \mathsf{H}\widehat{K}_\alpha\mathsf{H}^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}$$

$$\mathsf{H}\widehat{K}_\beta\mathsf{H}^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \qquad \mathsf{H}\widehat{K}_\gamma\mathsf{H}^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Note finally that H can be decomposed as a tensor product of two-dimensional forms

$$\mathsf{H} = \mathsf{h} \otimes \mathsf{h} \qquad \mathsf{h} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}. \tag{8}$$

Thus far, we have recovered the standard analysis, with the emergence of the Hadamard transformation as the key to resolving the Kimura model. For (multi)-taxon probability densities *evolving independently*, the time evolution after time t is by extension of (4)

$$P(t) = \mathrm{e}^{-\lambda t} \cdot \mathrm{e}^{\lambda t\widehat{\mathbb{T}}} \cdot P(0) \tag{9}$$

where P is a tensor of rank $\geqslant 2$ carrying the probability density on a sample space spanned by the appropriate Cartesian product of character sets, and $\widehat{\mathbb{T}} := \widehat{T} \otimes \mathbb{1} \otimes \cdots \mathbb{1} + \mathbb{1} \otimes \widehat{T} \cdots + \cdots$ the corresponding off-diagonal rate operator lifted to the tensor product space. Clearly, the higher rank Hadamard operator $\mathsf{H} \otimes \mathsf{H} \otimes \cdots$ again implements the correct diagonalization in this case. The work of [3–6] using discrete Fourier analysis on trees establishes that, remarkably, even when the multi-taxon system has evolved *via a phylogenetic tree*, the Hadamard transform technique *still* applies.

In order to pursue the alternative analysis via Lie symmetries, we exploit the observation that the Kimura operators $\widehat{K}_i$, $i \in \{\alpha, \beta, \gamma\}$ provide a Cartan subalgebra for transformations belonging to the Lie algebra $sl(4)$ of the group $SL(4)$ of (complex) matrices[6]. Clearly the Hadamard basis vectors $h_a := \mathsf{H} \cdot e_a$ are simultaneous eigenvectors of the Kimura generators. In a general representation of $SL(4)$, the eigenvalues of the Kimura generators simply

---

[6] $SL(4) \simeq GL(4)/\mathbb{C}^\times$ where invertible matrices are factored by the multiplicative group of complex numbers corresponding to their (nonzero) determinants.

correspond to the weight decomposition with respect to the Cartan subalgebra. Apart from the overall scaling by $e^{-\lambda t}$, the Markov transition operator $e^{-\lambda t} \cdot e^{\lambda t \widehat{\mathbb{T}}}$ appends an exponential time dependence given by the sum of these weights, multiplied by the Kimura 'charge' parameters α, β, γ. Thus, in principle, provided the Markov model respects the symmetry, its spectral properties, and hence the time development of a multi-taxon density, can be deduced from an appropriate weight decomposition of the corresponding tensor representation of $SL(4)$.

To confirm that the analysis does indeed carry through in the presence of phylogenetic trees, we now turn to the description of the branching process itself. The usual formalism of stochastic models of base substitution [8] can conveniently be encapsulated via a linear operator δ, which changes the state vector representing a single taxon, to that representing independent progeny after branching[7], defined by

$$\delta \cdot e_a = e_a \otimes e_a \qquad a = 1, \ldots, K. \tag{10}$$

For $K = 4$ in the nucleotide basis we have

$$\begin{aligned} \delta \cdot e_A = e_A \otimes e_A \qquad & \delta \cdot e_G = e_G \otimes e_G \\ \delta \cdot e_U = e_U \otimes e_U \qquad & \delta \cdot e_C = e_C \otimes e_C \end{aligned} \tag{11}$$

so that, when applied to a vector p representing the density on bases for one taxon, we have

$$\begin{aligned} p &= p_A e_A + p_G e_G + p_U e_U + p_C e_C \\ &\to \delta \cdot p = p_A e_A \otimes e_A + p_G e_G \otimes e_G + p_U e_U \otimes e_U + p_C e_C \otimes e_C \end{aligned} \tag{12}$$

after which evolution proceeds for the model on two taxa (with all operations lifted to the tensor product space carrying the Cartesian square of the character set, as described by (9) above).

A stochastic model may be said to possess a *symmetry* under a continuous transformation group G *if the rate matrix commutes with its generators*, and hence intertwines the group action. Formally if the action is $p(t) \to p'(t) \equiv g \cdot p(t)$ then the master equation (2) retains its form, for all $g \in G$ and for arbitrary $p(t)$, as

$$\frac{dp'(t)}{dt} = \widehat{R} \cdot p'(t)$$

iff $g\widehat{R} = \widehat{R}g$, or $[\widehat{R}, \widehat{K}] = 0$ with $\widehat{K}$ a *generator* of the group G ($g \sim e^{\widehat{K}}$). Similarly a branching operator δ admits a symmetry under such transformations if it intertwines the action of G on the character space of a single taxon, with some action on the tensor product space

$$\delta \circ g = \widetilde{g} \circ \delta. \tag{13}$$

Such symmetry considerations lead to useful ways of analysing the tree structure of general phylogenetic branching processes, which we hope to take up in a separate work. Here we examine the implications for the Kimura model as a first example. It is clear from the above remarks that the rate matrix admits a $GL(1) \times GL(1) \times GL(1) \simeq \mathbb{C}^{\times} \times \mathbb{C}^{\times} \times \mathbb{C}^{\times}$ group of symmetry tansformations. For purposes of weight labelling, and to make contact with standard group representation theory, it is convenient to consider the generators of the corresponding unitary phase subgroup $U(1) \times U(1) \times U(1)$, with the convention that group elements associated with compact generators within $SL(4)$ have pure imaginary parameters, whereas those with noncompact generators have real parameters. Turning to the diagonal branching operator (10), it is obvious that any symmetry generator acting as a permutation σ on the basic unit vectors e_a, $\sigma \cdot e_a = e_{\sigma a}$, will satisfy $\delta \circ \sigma \cdot e_a = e_{\sigma a} \otimes e_{\sigma a} = \sigma \otimes \sigma \circ \delta \cdot e_a$.

---

[7] A dynamical, many-body formulation of phylogenetic branching processes has been presented in [7].

In particular the Kimura generators in the distinguished basis do indeed permute the nucleotide unit vectors e_A, e_G, e_U, e_C and hence themselves have the diagonal intertwining property[8]:

$$\delta \circ \widehat{K}_i = \widehat{K}_i \otimes \widehat{K}_i \circ \delta \qquad i \in \{\alpha, \beta, \gamma\}. \tag{14}$$

Thus we conclude that the Kimura model has $U(1) \times U(1) \times U(1)$ symmetry, *both* in the sense of commuting with the rate matrix, *and* in the intertwining property for the branching operator.

With the above preliminaries we sketch briefly the way in which the above algebraic structure can be applied to an analysis of the Kimura model for phylogenetic trees, which is consistent with the Fourier transform methods. Fixing a rooted tree on L leaves, the full time evolution from the initial root density to the leaf density can be represented abstractly as a product of strings of terms of the form

$$\cdots (M_1' \otimes M_2' \otimes \cdots \otimes M_{r+1}') \cdot (\mathbb{1} \otimes \mathbb{1} \otimes \cdots \delta \otimes \cdots \otimes \mathbb{1}) \cdot (M_1 \otimes M_2 \otimes \cdots \otimes M_r) \cdots \tag{15}$$

where it is implied that, for the time slices $\Delta t, \Delta t'$ of the tree under consideration, with r taxa evolving, a branching event[9] took place on a particular edge leading to $r + 1$ taxa evolving, the M, M' being simply the appropriate Markov transition matrices $e^{\Delta t \widehat{R}}, e^{\Delta t' \widehat{R}'}$. The intertwining property (14) can now be used to pull all the δ operators back to the root node, so that the final expression for the leaf density is of the form of products of exponentials of tensor products of Kimura operators, acting on the fully branched state[10]

$$\delta^{(L-1)} p(0) = p_A(0) e_A \otimes e_A \cdots \otimes e_A + p_G(0) e_G \otimes e_G \cdots \otimes e_G$$
$$+ p_U(0) e_U \otimes e_U \cdots \otimes e_U + p_C(0) e_C \otimes e_C \cdots \otimes e_C. \tag{16}$$

Working in the Hadamard basis allows the exponentials to be diagonalized in terms of the weights of the tensor product states under the induced $U(1) \times U(1) \times U(1)$ action. The combinatorics of the tree is of course encoded, in that the change on each edge explicit in (15) is inherited by the differing total weights of each factor, and hence different exponential time dependence, in the L edges emanating from (16) above.

As an example we specialize to the binary character case (the symmetric two colour model [9, 10]). Suppose the character set is $\{Y, R\}$ for definiteness. The analogue of the Kimura operator is $\widehat{k}$, and there is only one rate parameter α with $\widehat{R} = \alpha(-\mathbb{1} + \widehat{k})$. The analogue of (8) is

$$\widehat{k} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \qquad h = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \qquad h\widehat{k}h^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}. \tag{17}$$

Consider the descending rooted 4-leaf tree $(1(2(34)))$ (see figure 1). Labelling the non-leaf edges 5, 6 in order of ascending level away from the leaves, define the total edge change parameters (including time intervals) as

$$\alpha_e := \Delta t_e \alpha \qquad e \in \{1, 2, \ldots, 6\}$$

(effectively allowing the α parameter in the rate matrix to be edge dependent), and also the leaf operators

$$\widehat{k}_1 = \widehat{k} \otimes \mathbb{1} \otimes \mathbb{1} \otimes \mathbb{1} \qquad \widehat{k}_2 = \mathbb{1} \otimes \widehat{k} \otimes \mathbb{1} \otimes \mathbb{1}$$
$$\widehat{k}_3 = \mathbb{1} \otimes \mathbb{1} \otimes \widehat{k} \otimes \mathbb{1} \qquad \widehat{k}_4 = \mathbb{1} \otimes \mathbb{1} \otimes \mathbb{1} \otimes \widehat{k}.$$

---

[8] In the case of Abelian algebras, a 'group-like' coproduct $\widehat{K} \to \widehat{K} \otimes \widehat{K}$ gives a coassociative coalgebra structure, and the tensor product spaces carry a consistent action.

[9] See also [7].

[10] The operator δ is coassociative, $(\mathbb{1} \otimes \delta) \circ \delta = (\delta \otimes \mathbb{1}) \circ \delta \equiv \delta^{(2)}$. Note that *no* off-diagonal terms such as $e_C \otimes e_U$ appear in the repeated application of δ to the initial state $p(0)$.
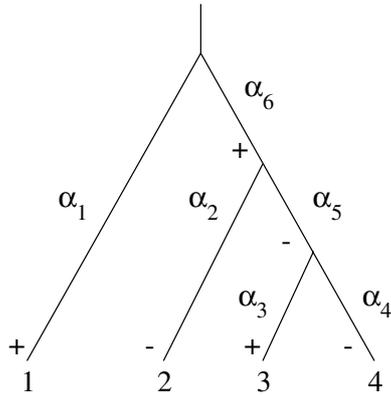
Figure 1. Descending 4-leaf tree (1(2(34)) with edges 1, 2, 3, 4, 5, 6 and leaf decoration $+, -, +, -$ indexing a component of P_{leaf} in the Hadamard basis. With signs propagated multiplicatively to the remaining edges, and with the overall $-\sum_e \alpha_e$ term, the exponent becomes $-2(\alpha_2 + \alpha_4 + \alpha_5)$, corresponding to the edge path sum for the split $\{2, 4\} \cup \{1, 3\}$.

Applying (14), (15), we have for the leaf density

$$P_{\text{leaf}} = \exp\left(-\sum_e \alpha_e\right) \exp(\alpha_1 \widehat{k}_1 + \alpha_2 \widehat{k}_2 + \alpha_3 \widehat{k}_3 + \alpha_4 \widehat{k}_4 + \alpha_5 \widehat{k}_5 + \alpha_6 \widehat{k}_6) \delta^3 p(0)$$

where also

$$\widehat{k}_5 = \mathbb{1} \otimes \mathbb{1} \otimes \widehat{k} \otimes \widehat{k} \qquad \widehat{k}_6 = \mathbb{1} \otimes \widehat{k} \otimes \widehat{k} \otimes \widehat{k}. \tag{18}$$

The composite operator in (18) acts in the Hadamard basis to give a signed sum of edge parameters, with the signs determined by products of $\widehat{k}$-weights, eigenvalues of the various leaf operators acting on $\delta^3 p(0) = p_Y(0) e_Y \otimes e_Y \otimes e_Y \otimes e_Y + p_R(0) e_R \otimes e_R \otimes e_R \otimes e_R$ expanded via the inverse Hadamard transform (see (17)),

$$e_Y = \tfrac{1}{2}(h_+ + h_-) \qquad e_R = \tfrac{1}{2}(h_+ - h_-). \tag{19}$$

Multiplying through by the overall prefactor, the *positively* signed edge parameters cancel in the exponent. For example, the coefficient of $h_+ \otimes h_- \otimes h_+ \otimes h_-$ in the expansion of (18) becomes

$$P_{+-+-} = \left(\tfrac{1}{2}\right)^4 e^{-2(\alpha_2+\alpha_4+\alpha_5)} p_Y(0) + \left(\tfrac{1}{2}\right)^2 \left(-\tfrac{1}{2}\right)^2 e^{-2(\alpha_2+\alpha_4+\alpha_5)} p_R(0).$$

As explained above, the use of (14), (15) in generalizing (18) to an arbitrary tree $\mathcal{T}$ amounts to considering how the symmetry group on the linear space spanned by the evolving probability density of a single system, extends after branching to transformations acting on the L-fold tensor product (in the Kimura 3ST model, the symmetry group is $U(1) \times U(1) \times U(1)$, and in the symmetric two-colour model just $U(1)$). Taking the binary case for simplicity, the general form of (18) reads (cf (9) and (4))

$$P_{\text{leaf}} = e^{-\sum_e \alpha_e} \cdot e^{\widehat{k}_{\mathcal{T}}} \cdot \delta^{(L-1)} p(0). \tag{20}$$

The operator $\widehat{k}_{\mathcal{T}}$ is essentially the induced generator of $U(1)$ after pulling back through the branching nodes of the tree. We define (following the above example)

$$\widehat{k}_{(e)} = \prod_{\ell \in \mathcal{T}_e} \widehat{k}_\ell$$

for each edge e to be the product, over all leaves in the subtree $\mathcal{T}_e$ determined by e, of the leaf operators

$$\widehat{k}_\ell = 1 \otimes 1 \otimes \cdots \widehat{k} \otimes 1 \otimes \cdots \otimes 1$$

($\widehat{k}$ acting on the ℓth place in the L-fold tensor product—obviously if ℓ is a leaf edge, $\widehat{k}_{(\ell)} \equiv \widehat{k}_\ell$). Then

$$\widehat{k}_\mathcal{T} = \sum_e \alpha_e \widehat{k}_{(e)}.$$

Finally, $\delta^{L-1} p(0)$ is the maximally branched state (as would derive from a multifurcating branching). Note however, that this decomposition does not imply that the leaf density is equivalent to independent stochastic evolution from this initial branched state—the operator $\widehat{k}_\mathcal{T}$ is not of separable form.

While (20) is basis independent, it is obviously beneficial to analyse the components of each side in terms of the (tensor products of) Hadamard vectors (eigenstates of the $\widehat{k}$ operator), as both the separable and non-separable parts of the tree operator $\widehat{k}_\mathcal{T}$ are diagonal in this basis. Briefly the algorithm for determining the weight attributed to a term of P_{leaf} in the Hadamard basis can be described as follows (see figure 1; for a formal analysis, see [12]). Take an arbitrary binary tree, and fix a tree 'split', to be associated with the coefficient, in the expansion of P_{leaf}, of the basis element consisting of the L-fold tensor product of $-$ Hadamard vectors on a chosen subset of distinguished leaves, with $+$ Hadamard vectors at the remaining non-distinguished leaf positions. On the graph of the tree assign $-$ signs to the distinguished leaf edges, and $+$ signs to the remainder, and propagate signs to the remaining edges multiplicatively (e.g. adjacent siblings with $-$ signs will generate a $+$ sign on their ancestral edge). The corresponding signed sum of edge parameters α_e is precisely the exponent generated by the action of $\widehat{k}_\mathcal{T}$ on this basis element. After the overall $\exp\left(-\sum_e \alpha_e\right)$ prefactor is multiplied through, *only* the negatively signed edge terms are present in the exponent (with coefficient -2). Finally the numerical factors accompanying the terms proportional to $p_Y(0)$ and $p_R(0)$ can easily be read off from (19).

It is clear that the above presentation is equivalent to the standard discrete Fourier analysis on tree techniques involving the Hadamard transform [3–6]. Specifically, the surviving edge parameters which provide the argument of the exponential are nothing but the nonintersecting path edge sums for a given leaf split, as emerges from the Hadamard transform in edge space. The standard $\mathbb{Z}_2 \times \mathbb{Z}_2$ colour symmetry is of course inherent in the Hadamard matrix, which is also mandatory for the simultaneous diagonalization of the Kimura generators. However, from the viewpoint of Lie symmetries, the latter determine three (infinite) continuous symmetry groups, rather than being identified with the three non-unit elements of a discrete group. Crucial for our derivation is the coproduct property (13), and the fact that the combinatorics of the tree determines the final action of the symmetry group on the L-fold tensor product carrying the leaf probability density.

In this letter we have provided a framework for the analysis of phylogenetic branching models on the basis of continuous transformation symmetries of the rate matrix and the branching operator. The formalism can be applied to the Kimura 3ST (and also the 2P) model, as well as the symmetric binary character model [9, 10] and it reproduces the standard spectral transform analysis. In summary, the main features of our work are as follows. (i) Given that the Fourier–Hadamard transform methods based on the Kinura models are tools for practical phylogenetic analysis for some datasets [5], it is important to supplement their established formal basis [4] with new derivations and insights. Besides generalizations (see below), the availability of explicit transform methods for nontrivial classes of model can be a significant analytical tool for rigorous examination of issues of tree reconstruction and phylogenetic

inference (see, for example, [11]). (ii) Our approach extends to any model where the off-diagonal rates can be associated with an Abelian subalgebra of $SL(K)$, whose generators have the form of permutation matrices (so that the intertwining property holds). For example, in place of (5)–(7) for $K = 4$ we could consider the (non-symmetric) $K = 3$ model

$$\widehat{\Gamma}_\alpha = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \qquad \widehat{\Gamma}_\beta = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

$$\widehat{T} = \frac{\alpha}{\alpha + \beta}\widehat{\Gamma}_\alpha + \frac{\beta}{\alpha + \beta}\widehat{\Gamma}_\beta \qquad \lambda = (\alpha + \beta). \tag{21}$$

The above analysis starting from (14) carries through, leading to an obvious analogue of (20) including both generators, and parameters α_e, β_e. The role of h is played by the 3×3 Fourier matrix

$$\mathsf{f} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & \omega & \omega^2 \\ 1 & \omega^2 & \omega \end{bmatrix} \qquad \omega^3 = 1 \tag{22}$$

and the eigenvalues of the tree operators $\widehat{\Gamma}_i^{\mathcal{T}}$, $i \in \{\alpha, \beta\}$ on P_{leaf} in the Fourier basis are computed starting from leaf decorations by cube roots of unity $1, \omega, \omega^2$ rather than ± 1 as in the binary split case. (iii) The branching operator δ defined in (10) is a central object whose implications for phylogenetic branching trees and network models deserve further study. (It is intimately related to the many-body, second-quantized formulation presented in [7]). (iv) In particular, objects such as $\delta^{(L-1)}p(0)$ or $\mathbb{1} \otimes \cdots \delta \otimes \cdots \otimes \mathbb{1} \circ P(t)$ possess special properties with respect to *entanglement* in the multi-taxon tensor space, whose characterization may provide powerful new tools for phylogenetic analysis.

We defer a formal presentation of further ramifications including generalizations, the role of Lie symmetries and representation theory in branching models in this context, and entanglement measures, to a separate work [12].

Acknowledgments

References

[1] Kimura M 1981 Estimation of evolutionary distances between homologous nucleotide sequences *Proc. Natl Acad. Sci. USA* **78** 454–8
[2] Semple C and Steel M 2003 *Phylogenetics (Oxford Lecture Series in Mathematics and Applications* vol 24) (Oxford: Oxford University Press)
[3] Steel M, Hendy M D and Penny D 1998 Reconstructing phylogenies from nucleotide pattern probabilities: a survey and some new results *Discr. Appl. Math.* **88** 367–96
[4] Székely L, Steel M A and Erdös P L 1993 Fourier calculus on evolutionary trees *Adv. Appl. Math.* **14** 200–16
[5] Hendy M D, Penny D and Steel M A 1994 Discrete Fourier analysis for evolutionary trees *Proc. Natl Acad. Sci. USA* **91** 3339–43
[6] Evans S N and Speed T P 1993 Invariants of some probability models used in phylogenetic inference *Ann. Stat.* **21** 355–77
[7] Jarvis P D and Bashford J D 2001 Quantum field theory and phylogenetic branching *J. Phys. A: Math. Gen.* **34** L703–7

[8] Rodriguez F, Oliver J L, Marin A and Medina J R 1990 The general stochastic model of nucleotide substitution *J. Theor. Biol.* **142** 485–501

[9] Farris J S 1973 A probability model for inferring evolutionary trees *Syst. Zool.* **22** 250–6

[10] Cavender J A 1978 Taxonomy with confidence *Math. Biosci.* **40** 271–80

[11] Chor B, Hendy M D, Holland B R and Penny D 2000 Multiple maxima of likelihood in phylogenetic trees: an analytic approach *Mol. Biol. Evol.* **17** 1529–1541

[12] Jarvis P D, Bashford J D and Sumner J G in preparation