MATRIX GROUP STRUCTURE AND MARKOV INVARIANTS IN THE STRAND SYMMETRIC PHYLOGENETIC SUBSTITUTION MODEL

PETER D. JARVIS AND JEREMY G. SUMNER

ABSTRACT. We consider the continuous time presentation of the strand symmetric phylogenetic substitution model (in which rate matrices are unchanged under Watson-Crick base conjugation). Analysis of the model’s underlying structure as a matrix group leads to a presentation of the rate generator matrix in an appropriate basis in terms of a two-part block decomposition. Markov invariants are classified and enumerated, and the quadratic invariants for the two-leaf phylogenetic tree are evaluated explicitly. Two of these Markov invariants are zero, and hence are concurrently phylogenetic invariants, while the other two invariants provide independent estimates of the total of substitution rates within the Watson-Crick conjugate pairs, and the total of remaining substitution rates across conjugate base pairs.

1. INTRODUCTION AND MOTIVATION

Recent years have seen rapid advances in the quantity and variety of molecular-based sequence data available for analysis and interpretation in terms of biological structure, function and evolution. Whole genome datasets are increasingly accompanied by other types of ‘–omic’ data: transcriptome, proteome, metabolome, amongst others. All of these modes of data representation in turn require adequate mathematical model building in stochastic settings, to capture the essential process systematics with parsimonious parametrizations.

Despite these ongoing challenges, the original brief of phylogenetics – the use of quantitative, interspecies comparison data (in the modern context, molecular sequence data), to infer the evolutionary ancestry of species – remains central. It is still the contention that quality data, based on suitably aligned molecular sequences, should admit analysis via appropriate parametric probability models consistent with the neutral theory of evolution. The aim is a statement of taxonomic ancestry via an inferred phylogenetic tree, or perhaps a network representation which encapsulates unresolved ambiguities in the data. In turn, under further assumptions about absolute mutation rates, parameter estimation then permits recovery of evolutionary divergence times.

For nucleic acid base sequence data, the so-called general Markov model is in practice specialized, so that the key theoretical object – an assumed $4 \times 4$ stochastic matrix of base substitutions – is not parametrized in the most general possible way. A popular choice for maximum likelihood calculations is the general time-reversible (GTR) model \[25\]; further constraints on the parameters lead to one of a number of other model types. Amongst these, we distinguish the so-called group-based models ([16], chapter 8), which allow for direct analytical treatments, using discrete Fourier or Hadamard inversion techniques \[7, 24\].

The armoury of theoretical techniques has been further enriched with the advent of geometrically-inspired methods which seek to locate certain algebraic varieties, defined by the embedding of the models’ parameter space into the multivariate probability spaces populated by the sequence data. Theoretical work around this approach is part of a new field of algebraic statistics \[15\]. The ideals
defining the polynomial varieties are simplest to characterize (via generating sets) in the group-based models [17], but have also been studied in the general Markov model [1].

Turning to computational approaches, although maximum likelihood optimization is powerful enough to allow full parameter recovery, in principle even for the general Markov model, in practical implementations it is usual to work with specialized models. In [19] we argued for the natural criterion of (multiplicative) closure as a guide to model choice in phylogenetics. In [19] it was shown that GTR generically fails to be multiplicative closed, and our subsequent work with simulations showed how serious errors in phylogenetic estimation could potentially arise as a result [23]. Beyond the group-based models, we have studied a large class of closed models based on matrix Lie groups, the so-called Lie Markov models [19]. In the continuous time case, their rate matrices belong to the affiliated Lie algebra, and the model itself is embedded within an appropriate stochastic cone (see [22] for details). The general Markov model itself is by construction closed, and in related work [18, 21] we have exploited its matrix group structure to construct many new polynomials in the probability tensor arrays which are group invariant – the so-called Markov invariants. These include, for example, for the quartet tree case, the remarkable ‘squangles’, degree five polynomials which act as powerful quartet identifiers for the general Markov model, without the need for full parameter reconstruction [21, 9].

Recently, in a study of rodent phylogeny, a hitherto un-noticed interesting regime of substitution parameters was pointed out by Yap and Pachter [27]. They identified in their analysis, a special case of the GTR parameters, wherein the substitution matrix becomes invariant under Watson-Crick base conjugation (in consequence, the stationary base frequencies also satisfy \( \pi_A = \pi_T, \pi_C = \pi_G \), consistent with Chargaff’s rule). This model class was formally introduced as the ‘strand-symmetric’ model, and its defining ideals in the algebraic geometry approach considered in detail, by Casanellas and Sullivant [3].

The purpose of this note is to point out that the strand symmetric model indeed has the closure property, and to exploit this in the continuous time case, to identify the corresponding matrix Lie algebra, and accompanying stochastic cone from which valid rate matrices must be constructed. In §2 below, we review the formal definition of the strand symmetric model in the context of phylogenetic substitution models. A decomposition of its Lie algebra in terms of the classical groups [26] is then given.

In §3 we conclude with some applications using the decomposition to enumerate related polynomial modules, allowing for the identification of both Markov invariants, and also, in principle, phylogenetic invariants at low degree. In particular, for the two-leaf phylogenetic tree, we construct four quadratic Markov invariants, two of these turn out to simultaneously be phylogenetic invariants, while the remaining two provide a method for independently estimating the total of substitution rates within, and across, Watson-Crick conjugate base pairs.

ACKNOWLEDGEMENT

Part of this work was completed under an Australian senior Fulbright scholarship (Department of Statistics, University of California Berkeley, and Department of Physics, University of Texas at Austin) and hosts and colleagues at these institutions are thanked for their support. JGS was partially supported by Australian Research Council grants DP0877447, FT100100031, and DE130100423.

2. THE STRAND SYMMETRIC RATE MODEL AND ITS LIE ALGEBRA STRUCTURE

The central construct in the standard theoretical approach to phylogenetic branching is an assumed substitution matrix parametrizing the probabilities for transitions between different states of a random
variable which encodes the stochastic nature of biological molecular sequences (bases, for nucleic acids, or amino acids, for proteins). Concentrating on DNA, we have for example a 2 state system \{R, Y\}, or a 4 state system with state space \{A, C, G, T\}.

Consider firstly the two state case. The general Markov model in this case has substitution matrix

\[
M = \begin{pmatrix}
m_{RR} & m_{RY} \\
m_{YR} & m_{YY}
\end{pmatrix}.
\]

Probability conservation constrains each row of \(M\) to have unit sum, so that there are two independent parameters \(m_{RY} \equiv a, m_{YR} \equiv b\), with \(M\) in the form

\[
M(a, b) = \begin{pmatrix}
1 - a & a \\
b & 1 - b
\end{pmatrix},
\]

Noting the closure property given by the matrix multiplication rule

\[M(a, b)M(a', b') = M(a(1 - a' - b') + a', b(1 - a' - b') + b'),\]

we therefore characterize the general two-state Markov model as the set of substitution matrices \(M(a, b)\) with \(0 \leq a, b \leq 1\). In order to apply group-theoretic methods, we enlarge the set \(M(a, b)\) by working over the complex field and removing any constraints other than \(\det(M(a, b)) = 1 - a - b \neq 0\), thereby defining a certain matrix subgroup of the general linear group of nonsingular 2\times2 matrices. In the usual way, this group possesses a Lie algebra, its tangent space at the identity defined via derivatives and generated in this case by \(R_1 := (\partial/\partial a)M(a, b)|_{a=b=0}, R_2 := (\partial/\partial b)M(a, b)|_{a=b=0}\), namely

\[
R_1 = \begin{pmatrix}
\overline{T} & 1 \\
0 & 0
\end{pmatrix}, \quad R_2 = \begin{pmatrix}
0 & 0 \\
1 & T
\end{pmatrix},
\]

with the only non-trivial commutator bracket given by \([R_1, R_2] := R_1R_2 - R_2R_1 = -R_1 + R_2\).

It is a general fact that, for arbitrary complex combinations \(Q = \alpha R_1 + \beta R_2\) in the Lie algebra, the matrix exponential \(\exp(Q)\) belongs to the corresponding matrix group. In order to recover the Markov substitution model however, the off-diagonal matrix elements of such \(Q\) should be positive quantities interpretable as substitution rates for the respective state transitions. Adopting a uniform normalization to negative unit trace, we characterize the two state Markov rate model as the set of matrices \(M = \exp(tQ), t > 0\), with \(Q = \alpha R_1 + \beta R_2\) and \(\alpha, \beta \geq 0, \alpha + \beta = 1\).

The situation for the general Markov substitution and rate models for the 4 state system, with state space \{A, C, G, T\}, is similar. Allowing for the row sum constraint, the Markov matrix

\[
M = \begin{pmatrix}
m_{AA} & m_{AC} & m_{AG} & m_{AT} \\
m_{CA} & m_{CC} & m_{CG} & m_{CT} \\
m_{GA} & m_{GC} & m_{GG} & m_{GT} \\
m_{TA} & m_{TC} & m_{TG} & m_{TT}
\end{pmatrix}
\]

has 12 free parameters, and the corresponding matrix group has Lie algebra spanned by \(6 + 6 = 12\) standard generators analogous to \(R_1, R_2\) above (two sets of six with positive unit entries above and below the diagonal, respectively, each with corresponding diagonal \(-1\)'s). The general Markov rate model consists therefore of convex combinations elements of the Lie algebra in the above basis (with nonnegative real coefficients), thus having negative unit trace. Our interest here is in restricted model classes having the closure property, and the affiliated matrix subgroups of the general Markov model.

The rate model for such a restricted class is then the intersection of the Lie subalgebra in question,

---

1Technically a matrix semigroup.
2For aesthetic purposes here and below, signed entries in matrices are written with overbars.
with the general Markov rate model as above. We refer to these rate matrices as the stochastic cone of the Lie algebra[4].

Consider now the general time reversible (GTR) model, where the guiding assumption is that transition rates involving arbitrary states \( i, j \in \{A, C, G, T\} \), weighted by the (stationary) distribution of the starting state \( \pi_k \), are independent of whether the transition is from \( i \) to \( j \), or \( j \) to \( i \), technically stated as

\[
\pi_i Q_{ij} = \pi_j Q_{ji}.
\]

In practice, this is implemented by taking an arbitrary symmetric matrix \( S \), and forming the (off diagonal) parts of \( Q \) as the product of \( S \) with the diagonal matrix of the stationary distribution,

\[
Q = \begin{pmatrix}
Q_{AA} & S_{AC} \pi_C & S_{AG} \pi_G & S_{AT} \pi_T \\
S_{CA} \pi_A & Q_{CC} & S_{CG} \pi_G & S_{CT} \pi_T \\
S_{GA} \pi_A & S_{GC} \pi_C & Q_{GG} & S_{GT} \pi_T \\
S_{TA} \pi_A & S_{TC} \pi_C & S_{TG} \pi_G & Q_{TT}
\end{pmatrix},
\]

with \( S_{ij} = S_{ji} \) and the diagonal entries set to ensure probability conservation (zero row sums for rate matrices), for example \( Q_{AA} = -S_{AC} \pi_C - S_{AG} \pi_G - S_{AT} \pi_T \). As required, the row vector of stationary probabilities \( (\pi_A, \pi_C, \pi_G, \pi_T) \) is a left null eigenvector of \( Q \).

A special case of the GTR model occurs when its transition rates are unchanged under Watson-Crick base pairing conjugation (i.e. \( A \leftrightarrow T, C \leftrightarrow G \)); for example \( Q_{CA} = Q_{GT}, Q_{CT} = Q_{GA}, Q_{TA} = Q_{AT} \), and so on. In the above parametrization, imposition of this constraint on self-conjugate pairs such as \( Q_{TA} = Q_{AT} \) enforces Chargaff’s rule on the stationary distribution, \( \pi_A = \pi_T \), and \( \pi_C = \pi_G \), and the remaining conditions constrain \( S \) also to fulfil the analogous conditions \( S_{AC} = S_{GT}, S_{AG} = S_{CT} \) etc. (for self-conjugate pairs, the relations \( S_{GG} = S_{GC} \) and \( S_{AT} = S_{TA} \) are already enforced by the symmetry of \( S \)). As mentioned, the GTR model class is not multiplicatively closed[19], and neither will this base pairing conjugation symmetric case be. Remarkably however, the strand symmetric model defined to fulfil the base pairing conjugation symmetry condition, does have the closure property, as follows.

A convenient parametrization of the strand symmetric model occurs by fixing an arbitrary minimal set of transition probabilities, and duplicating these entries in the conjugate matrix elements. Thus we choose

\[
M = \begin{pmatrix}
m_{AA} & m_{AC} & m_{AG} & m_{AT} \\
m_{CA} & m_{CC} & m_{CG} & m_{CT} \\
m_{GA} & m_{GC} & m_{GG} & m_{GT} \\
m_{TA} & m_{TC} & m_{TG} & m_{TT}
\end{pmatrix} \equiv \begin{pmatrix}a & b & c & d \\
e & f & g & h \\
h & g & f & e \\
d & c & b & a
\end{pmatrix}
\]

where \( a \equiv 1 - b - c - d, f \equiv 1 - e - g - h \). That closure indeed holds, follows trivially by verifying that the matrix product \( MM' \) of two such patterned matrices respects the base conjugation symmetry.

In terms of the model classes referred to in the introductory discussion, the strand symmetric model occurs as an ‘equivariant model’ [4], which are a useful generalisation of the standard ‘group-based’ models ([16], chapter 8) and are multiplicatively closed. Other examples are the Kimura three parameter model with \( b = e, c = h, g = d \), the Kimura two parameter model with \( b = e = g = d, c = h \), and the Jukes-Cantor (one parameter) model with \( b = c = h = e = g = d \). In particular, the strand symmetric model is constructed as an equivariant model by including all substitution matrices \( M \) invariant under simultaneous row and column permutations drawn from \( \{\epsilon, (AT)(GC)\} \) (where \( \epsilon \) is

---

[1] Consult [22]; details of the general case are not required in the present work.
the identity or ‘do nothing’ permutation), as is clear from the explicit form given above. As noted earlier, a broader approach to multiplicatively closed model classes, where the state space of the Markov chain is deemed to have some structure invariant under a fixed group of state permutations, has been presented in [19, 22] under the banner of ‘Lie-Markov’ models. In that work, a somewhat broader notion of model symmetry is utilized, where a model is deemed to have a certain permutation symmetry, not if each individual substitution matrix is invariant under permutations drawn from the group (as in the equivariant case), but rather if each permutation produces a (possibly distinct) substitution matrix which is also included in the model. This notion of symmetry allows for permutations of individual parameters in the model, which, as is argued in [19], is consistent with the fact that the parameter labels play no intrinsic role, as parameters must be fitted to data using statistical inference. In particular, in [22] a complete hierarchy consisting of 35 multiplicatively closed models is derived, which are additionally invariant under the permutations which fix the partitioning of nucleotides into purines and pyrimidines, i.e. \( AG|CT := \{\{A, G\}, \{C, T\}\} \), so notationally \( AC|GT \equiv GA|CT \equiv TC|AG \) etc. In [22] it is also noted that an equivalent hierarchy exists for the partitioning that defines the Watson-Crick base pairing conjugation, i.e. \( AT|GC \) (and yet another hierarchy for the partitioning \( AC|GT \)). In particular, Model 6.6 [22] is identical to the strand symmetric model with the substitution \( G \leftrightarrow T \) (or \( A \leftrightarrow C \)). From this point of view, the strand symmetric model lives in a large hierarchy of Lie Markov models, equivalent to the hierarchy presented in [22], where each model has symmetry consistent with Watson-Crick base pairing.

Let \( S \) be the vector space associated with the four nucleotide bases, with standard unit vectors \( e_A = (1, 0, 0, 0), \ e_C = (0, 1, 0, 0), \ e_G = (0, 0, 1, 0), \ e_T = (0, 0, 0, 1) \), so \( S := \{e_A, e_C, e_G, e_T\}_C \cong \mathbb{C}^4 \) and, for example, the stationary distribution is given by the vector \( \pi = \pi_A e_A + \pi_C e_C + \pi_G e_G + \pi_T e_T \). Following the analysis in the two state case, we consider the matrix Lie group affiliated to the strand symmetric model. In the usual way of extracting the Lie algebra as the tangent space at the identity, we find, in direct correspondence with variations in the independent parameters \( b, c, d, e, g, h \), the following six generators:

\[
S_1 = \begin{pmatrix}
\bar{1} & 1 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 1 & \bar{1}
\end{pmatrix}, \quad
S_2 = \begin{pmatrix}
\bar{1} & 0 & 1 & 0 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 1 & 0 & \bar{1}
\end{pmatrix}, \quad
S_3 = \begin{pmatrix}
\bar{1} & 0 & 0 & 1 \\
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
1 & 0 & 0 & \bar{1}
\end{pmatrix},
\]

\[
T_1 = \begin{pmatrix}
0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 \\
0 & 0 & 1 & 1 \\
0 & 0 & 0 & 0
\end{pmatrix}, \quad
T_2 = \begin{pmatrix}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 \\
0 & 1 & 1 & 0
\end{pmatrix}, \quad
T_3 = \begin{pmatrix}
0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 \\
0 & 1 & \bar{1} & 0
\end{pmatrix}.
\]

In this way, we can represent a rate matrix \( Q \) as

\[
Q = \alpha_1 S_1 + \alpha_2 S_2 + \alpha_3 S_3 + \beta_1 T_1 + \beta_2 T_2 + \beta_3 T_3,
\]

where \( \alpha_1, \alpha_2, \alpha_3 \) and \( \beta_1, \beta_2, \beta_3 \) are generic parameters. Moreover, it is easily checked that the \( \text{Ansatz} \) \( \pi_A = \pi_T = p, \pi_C = \pi_G = q \) for \( (\pi_A, \pi_C, \pi_G, \pi_T) \), provides a left null eigenvector of the transition matrix \( Q \), if \( p = (\beta_1 + \beta_2)/2(\alpha_1 + \alpha_2 + \beta_1 + \beta_2), \ q = (\alpha_1 + \alpha_2)/2(\alpha_1 + \alpha_2 + \beta_1 + \beta_2) \) – independently of \( \alpha_3 \) and \( \beta_3 \) – which is therefore the unique stationary distribution. A graphical representation of the model is given in Figure[1].

---

4The exact number of models in the hierarchy depends somewhat on whether certain special cases are included in the count or not. The complete hierarchy, together with full details of matrix elements for each model, is provided online at www.pagines.mal.upc.edu/~jfernandez/LMNR.pdf.
The full set of 15 commutation relations amongst these generators is
\[
\begin{align*}
[S_1, S_2] &= S_1 - S_2, & [S_2, S_3] &= -S_1 + S_2, & [S_3, S_1] &= -S_1 + S_2, \\
[T_1, T_2] &= T_1 - T_2, & [T_2, T_3] &= -T_1 + T_2, & [T_3, T_1] &= -T_1 + T_2, \\
[S_1, T_1] &= -S_1 + T_1, & [S_1, T_2] &= -S_1 + S_3 - T_3 + T_2, & [S_1, T_3] &= -S_1 + S_2, \\
[S_2, T_1] &= -S_2 + S_3 + T_1 - T_3, & [S_2, T_2] &= -S_2 + T_2, & [S_2, T_3] &= S_1 - S_2, \\
[S_3, T_1] &= T_1 - T_2, & [S_3, T_2] &= -T_1 + T_2, & [S_3, T_3] &= 0,
\end{align*}
\]
as can be checked by elementary matrix algebra. We denote the corresponding Lie algebra by \( l_{SSM} \).

The permutations
\[
\{\epsilon, (AT), (GC), (AT)(GC), (AG)(CT), (AC)(GT), (AGTC), (ATGC)\},
\]
fix the Watson-Crick pairing \( AT|GC \), and are generated, for example, by the permutation \( (AT) \), via \( TA|GC \equiv AT|GC \), and the permutation \( (AG)(TC) \), via \( GC|AT \equiv AT|GC \). In terms of the generators of the Lie algebra \( l_{SSM} \), these permutations produce the label substitutions \( 1 \leftrightarrow 2 \) and \( S \leftrightarrow T \), respectively.

We now proceed via Levi’s theorem \([5]\) to give the structure of \( l_{SSM} \) as the direct sum of a semisimple and a solvable part using the following matrix notation. We denote the unique three-dimensional simple Lie algebra \( A_1 \cong B_1 \cong C_1 \) in Cartan’s classification as \( sl_2 \), and the one-dimensional (abelian) Lie algebra \( \cong \mathbb{C} \) as \( gl_1 \). As generators of the so-called ‘defining’ representation of \( sl_2 \), with corresponding module \( U \cong \mathbb{C}^2 \), we take:
\[
K_+ = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad K_- = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad K_0 = \frac{1}{2} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},
\]
with commutation relations \([K_0, K_\pm] = \pm K_\pm, [K_+, K_-] = 2K_0\). The Lie algebra affiliated to the two-dimensional general Markov model, with generators \( R_1 \) and \( R_2 \) described above, is isomorphic to the unique nonabelian two-dimensional Lie algebra \([5]\), consisting of the semidirect sum of a one-dimensional abelian algebra with a one-dimensional factor (often referred to as the ‘shift algebra’), generated by \( X \) and \( Y \) with non-zero commutation relation \([X, Y] = Y\). We denote this Lie
algebra by \( l_2 \) and consider the representation\(^5\) obtained by taking the module \( \mathcal{V} \cong \mathbb{C}^2 \) and generators \( X = \frac{1}{2} (R_1 + R_2) \) and \( Y = \frac{1}{2} (R_2 - R_1) \), i.e.

\[
X = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad Y = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.
\]

Note any module \( \mathcal{W} \) of \( sl_2 \oplus l_2 \) extends to a module \( \mathcal{W}_r := \mathcal{W} \otimes \mathcal{R} \cong \mathcal{W} \) of \( sl_2 \oplus l_2 \oplus gl_1 \), where \( \mathcal{R} \cong \mathbb{C} \cong \langle \nu \rangle \) and \( \nu \) is an eigenvector of a generator \( R \) of \( gl_1 \). Below we will also have recourse to refer to the “trivial” representations of \( sl_2 \) and \( gl_1 \), with modules \( \mathcal{U}_0 \cong \mathbb{C} \) and \( \mathcal{R}_0 \cong \mathbb{C} \) respectively, obtained by mapping all generators to 0. We also require an additional representation of \( l_2 \) with module \( \mathcal{V}' \cong \mathbb{C} \) where \( \hat{X} \) is mapped to a fixed square matrix (of arbitrary size), \( Y \) is mapped to 0, and \( \mathcal{V}' = \langle \nu' \rangle \cong \mathbb{C} \) with \( \nu' \) an eigenvector for the image of \( X \).

**Lemma 1: Decomposition of the Lie algebra of the strand symmetric model**

The Lie algebra \( l_{SSM} \) generated by \( S_1, S_2, S_3, T_1, T_2 \) is isomorphic to the direct sum \( sl_2 \oplus gl_1 \oplus l_2 \) of the simple three-dimensional Lie algebra \( sl_2 \), a one-dimensional Lie algebra \( gl_1 \), and the two-dimensional shift algebra \( l_2 \).

**Proof:** Define the new set of generators,

\[
\hat{K}_0 = \frac{1}{4} (-S_3 + T_3), \quad \hat{K}_+ = \frac{1}{2} (S_1 - S_2), \quad \hat{K}_- = \frac{1}{2} (T_1 - T_2);
\]

\[
\hat{R} = \frac{1}{2} (S_3 + T_3);
\]

\[
\hat{X} = \frac{1}{4} (S_1 + S_2 + T_1 + T_2), \quad \hat{Y} = \frac{1}{4} (-S_1 - S_2 + S_3 + T_1 - T_3 + T_2).
\]

By direct computation,

\[
\hat{K}_0 = \frac{1}{4} \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}, \quad \hat{K}_+ = \frac{1}{2} \begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}, \quad \hat{K}_- = \frac{1}{2} \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix};
\]

\[
\hat{R} = \frac{1}{2} \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}; \quad \hat{X} = \frac{1}{4} \begin{pmatrix} 2 & 1 & 1 & 0 \\ 1 & 2 & 0 & 1 \\ 1 & 0 & 2 & 1 \\ 0 & 1 & 1 & 2 \end{pmatrix}, \quad \hat{Y} = \frac{1}{4} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix},
\]

we find the only non-zero commutation relations are \([\hat{K}_0, \hat{K}_\pm] = \pm \hat{K}_\pm, [\hat{K}_+, \hat{K}_-] = 2\hat{K}_0, [\hat{X}, \hat{Y}] = \hat{Y}\), as required.

\[\square\]

**Lemma 2: Decomposition of the state space \( \mathcal{S} \) of strand symmetric model**

As a module of \( l_{SSM} \cong sl_2 \oplus gl_1 \oplus l_2 \), the four-dimensional state space \( \mathcal{S} \) decomposes as the direct sum of two two-dimensional components \( \mathcal{S} = U \oplus V \) where

\[
U \cong (\mathcal{U} \otimes \mathcal{R} \otimes \mathcal{V}') ,
\]

\[
V \cong (\mathcal{U}_0 \otimes \mathcal{R}_0 \otimes \mathcal{V}) ,
\]

and

1. \( \mathcal{U} \cong \mathbb{C}^2 \) and \( \mathcal{U}_0 \cong \mathbb{C}^1 \) are the \( sl_2 \) modules described above,

---

\(^5\)See [20] for an algebraic investigation of the role of \( l_2 \) in phylogenetic tree and network models.
(2) $\mathcal{R} \cong \mathbb{C}$ and $\mathcal{R}_0 \cong \mathbb{C}$ are the $gl_1$ modules described above,
(3) $\mathcal{V} \cong \mathbb{C}^2$ and $\mathcal{V}' \cong \mathbb{C}$ are the $l_2$ modules described above.

Proof:

As an alternative ordered basis for $\mathcal{S}$, take $\{u_0, u_1, v_0', v_1'\}$ where
\[
 u_0 = (1, 0, 0, 1), \quad u_1 = (0, 1, 1, 0), \quad v_0' = (1, 0, 0, 1), \quad v_1' = (0, 1, 1, 0).
\]

By direct computation, taking $\{u_0, u_1, v_0', v_1'\}$ as an ordered basis, we have the block forms (where boldface $1$ denotes the $2 \times 2$ identity matrix):
\[
 \hat{K}_0 = \begin{pmatrix} K_0 & 0 \\ 0 & 0 \end{pmatrix}, \quad \hat{K}_+ = \begin{pmatrix} K_+ & 0 \\ 0 & 0 \end{pmatrix}, \quad \hat{K}_- = \begin{pmatrix} K_- & 0 \\ 0 & 0 \end{pmatrix},
\]
\[
 \hat{R} = \begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \hat{X} = \begin{pmatrix} -\frac{1}{2} \textbf{1} & 0 \\ 0 & X \end{pmatrix}, \quad \hat{Y} = \begin{pmatrix} 0 & 0 \\ 0 & Y \end{pmatrix}.
\]

Inspection of the blocks completes the proof. □

In the applications below we will refer to $\{u_0, u_1, v_0', v_1'\}$ as the ‘split’ basis of $\mathcal{S}$.

3. Discussion and Applications

Our aim thus far has been to present the strand symmetric model $[27, 3]$ from the point of view of the underlying continuous Lie group. In Lemma 1 we gave a classical classification of the Lie algebra associated with the strand symmetric model, and established the remarkable block diagonal form presented in Lemma 2. We now turn to simple applications of these results.

In the analysis of [3], the self-similar structure of its substitution matrix was exploited to formulate the strand symmetric model as a generalization from group-based models to matrix valued group based models. This allows known Fourier Hadamard inversion techniques to be pursued, and in different situations, the ideal structure of the appropriate algebraic varieties can be described (including generalizations of the linear invariants, well-known from the vanishing coefficients in the Hadamard basis occurring in the standard group-based models).

Our approach with Lie group methods provides complementary insights. From the point of view of distance measures for phylogenetic reconstruction, any model can be subjected to tools such as the LogDet $[2, 12, 14]$. The LogDet arises as a particular example of the more general concept of Markov invariants $[18]$, which are polynomials providing one-dimensional representations of the Lie group underlying a given phylogenetic model. However, the great numerical appeal of the linear inversions, that the Hadamard conjugation provides for the Kimura three parameter model $[8]$, is the availability of phylogenetic information via nothing more than a change of basis. This should be compared to the polynomial calculations (as required by Markov invariants when the underlying Lie group arises from the general Markov model of sequence evolution), which are inherently more susceptible to stochastic error. In the case of Markov models with additional special symmetries, such as the strand symmetric model, it is of significant benefit that lower degree Markov invariants provide equivalent information to the LogDet (which, for a state space of size four such as DNA, is a degree 4 polynomial). As the matrix Lie group underlying the strand symmetric model is nonabelian (as exhibited by non-zero commutation relations in $l_\text{SSM}$), a complete set of linear invariants is not available (and hence no linear inversion technique analogous to the Hadamard conjugation is applicable);

\[\text{Here and below we will mark the indices of vectors (and/or tensor components) in } V \text{ by }'.\]
however, it turns out that a hierarchy of quadratic Markov invariants can be deployed for any number of leaves. The following is derived in the appendix, Appendix A, which uses our previously established rules for working out the appropriate representations. Here we quote the main result:

**Lemma 3: Quadratic Markov invariants for the strand symmetric model**

1. For \( L \) leaves there are precisely \( \frac{1}{2} (3^L + (-1)^L) \) linearly independent quadratic Markov invariants for the strand symmetric model, namely 1, 5, 13, 41, \( \cdots \) for \( L = 1, 2, 3, 4, \cdots \) respectively.

2. For \( L = 2 \) case for two-leaf trees (or cherries, from marginalizations of larger alignments) the count of 5 includes the square of the total probability mass (the linear invariant), as well as two phylogenetic invariants which vanish identically. These and the two remaining nonzero Markov invariants \( I_1, \cdots \) are derived in Appendix A below.

**Proof:**

See Lemma 2 above, and Appendix A below.

In the 2 leaf case the quadratic invariants are proxies for determinant functions, not of the full \( 4 \times 4 \) probability array, but for its \( 2 \times 2 \) blocks in the split basis provided by the decomposition of the state space given in Lemma 2, and as such can provide differential information about the relative contributions of the rate parameters to the total edge lengths. In particular, these invariants provide a method for estimating the total sum of rates \( (\alpha_3 + \beta_3) \) multiplied by time elapsed within the Watson-Crick conjugate pairs, and the total sum of rates \( (\alpha_1 + \beta_1 + \alpha_2 + \beta_2) \) multiplied by time elapsed across the Watson-Crick conjugate pairs (refer to Figure 1 for illustration). This is realized in the explicit constructions in Appendix A and should be compared to application of the LogDet, which confounds these two quantities into a total sum.

**APPENDIX A. MARKOV INVARIANTS FOR THE STRAND SYMMETRIC MODEL**

In recent work we have studied the enumeration and applicability of certain one-dimensional representations of Markov-type groups, which we have termed Markov invariants. The following discussion adopts the notation and results of [18, 21, 10], and especially [11].

In the language of representation theory, polynomials in phylogenetic pattern tensors are technically polynomial representations of the underlying matrix groups. For general matrix groups, the starting point is the representations of the general linear group \( GL(n) \) or equivalently its Lie algebra \( gl_n \), where the irreducible representations are labelled by (ordered) integer partitions \( \lambda \) of \( m \), and we write \( \lambda \vdash m \) and \( \lambda = (\lambda_1, \lambda_2, \ldots, \lambda_r) \), with \( \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r \geq 0 \) and \( \sum \lambda_i = m \). When a partition \( \lambda \) refers directly (and equivalently) to a specific irreducible representation, module, or character of \( gl_n \), we adopt Littlewood’s notation for Schur functions, where the partition is enclosed by curly brackets: \( \{ \lambda \} \).

For a module \( S \) of a matrix group \( G \leq GL(n) \), polynomials of degree \( D \) in the components of \( S \) belong to the (in general reducible) module \( S \otimes \{ D \} \), the plethysm of \( S \) with the one-part partition \( \lambda = (D) \). For an \( L \)-way phylogenetic pattern tensor, the module is the \( L \)-fold tensor product of the corresponding product group \( G \times G \times \cdots \times G \) (one copy for each pendant edge). In this case the resolution of \( (\otimes^L S) \otimes \{ D \} \) requires calculation of generic plethysms \( S \otimes \sigma \), where \( \sigma \vdash D \). Further, the multiplicities \( g^\lambda_{\mu \nu} \), with \( \lambda, \mu, \nu \vdash D \), which resolve tensor products (inner multiplication ‘\(*\)’) of irreducible modules in the symmetric group \( \mathfrak{S}_D \), must also be computed. The following result is

---

7 Integer sequence A046717 (see http://oeis.org/).
Lemma 4: Markov invariants:

To enumerate Markov invariants at degree \( D \), carry out the following steps:

1. For each \( \sigma \vdash D \), compute the number of one-dimensional representations \( f_\sigma \) occurring in the decomposition of \( S \otimes \{ \sigma \} \).

2. The number of Markov invariants at degree \( D \) is then

\[
\begin{aligned}
n_D = \sum_{\sigma_1, \sigma_2, \ldots, \sigma_L \vdash D} g_{\sigma_1 \sigma_2 \ldots \sigma_L}^{(D)} f_{\sigma_1} f_{\sigma_2} \cdots f_{\sigma_L}
\end{aligned}
\]

where \( g_{\sigma_1 \sigma_2 \ldots \sigma_L}^{(D)} \) is the inner product multiplicity for the occurrence of the module \( (D) \) in the tensor product \( \sigma_1 \otimes \sigma_2 \otimes \cdots \otimes \sigma_L \) of modules of \( \mathfrak{S}_D \).

As a simple first case, consider \( D = 2 \). The partitions \( \sigma \vdash D \) are simply \( S := (2) \) and \( A := (1^2) \), and we adopt these symbols for the corresponding characters. In \( \mathfrak{S}_2 \) the inner products \( S * S = S \), \( A * S = A \), \( A * A = S \) are well known and so \( g_{S \ast S}^{S} = 1 \), \( g_{S \ast A}^{S} = 0 \), \( g_{S \ast A}^{A} = 1 \) and in general the inner product multiplicity is 1 or 0 for an even or odd number of \( A \)'s, respectively. However, as will be seen shortly, \( f_A = 2 \), and \( f_S = 1 \). Hence applying Lemma 4, \( n_2 = \sum_{\ell=0}^{\lfloor L/2 \rfloor} \binom{L}{2\ell} 2^{2\ell} \) which yields the formula given in Lemma 3 above. Note that the single, linear Markov invariant is simply the probability mass, normalized to 1.

Lemma 5: Calculation of \( f_\sigma \) for \( S \) at degree \( D = 2 \)

Proof: We appeal to the left-distributive law for plethysms [13]:

\[
(A + B) \otimes C = \sum_{\mu \in C} (A \otimes C/\mu) \otimes (B \otimes \mu)
\]

where \( A, B, C \) are \( \mathfrak{gl}_n \) characters and the summation is over all \( \mu \) where the skew character \( C/\mu \) is defined. Referring to Lemma 2 and ignoring the modules \( \mathcal{R}, \mathcal{R}_0, \mathcal{V}' \) and \( \mathcal{U}_0 \), which being one-dimensional do not influence our calculation, we take \( A \equiv \mathcal{U} \) as a \( \mathfrak{sl}_2 \) \( < \mathfrak{gl}_2 \) module and \( B \equiv \mathcal{V} \) as a \( \mathfrak{sl}_2 \) \( < \mathfrak{gl}_2 \) module. We then compute

\[
(A + B) \otimes \{2\} = \sum_{\mu = \{0\}, \{1\}, \{2\}} (A \otimes \{2\}/\mu) \otimes (B \otimes \mu)
\]

\[
= (A \otimes \{2\}) \otimes (B \otimes \{0\}) + (A \otimes \{1\}) \otimes (B \otimes \{1\}) + (A \otimes \{0\}) \otimes (B \otimes \{2\})
\]

\[
= A \otimes \{2\} + A \otimes B + B \otimes \{2\}
\]

where, in the final line, we have implemented the plethysms \( A \otimes \{1\} = A \) and \( B \otimes \{1\} = B \), and removed the trivial plethysms \( A \otimes \{0\} \) and \( B \otimes \{0\} \) (which, incidentally, correspond exactly to the modules \( \mathcal{U}_0 \cong \mathbb{C} \) and \( \mathcal{V}_0 \cong \mathbb{C} \), respectively). Now, considered as an \( \mathfrak{sl}_2 \) module, \( A \otimes \{2\} \) is irreducible with dimension 3, and similarly considered as a \( \mathfrak{sl}_2 + \mathfrak{l}_2 \) module \( A \otimes B \) has dimension \( 2 \times 2 = 4 \) and is irreducible because \( A \) is irreducible. However, the general theory in [18] establishes that \( B \otimes \{2\} \) contains a one-dimensional submodule of \( \mathfrak{l}_2 \); hence we conclude \( f_S = 1 \). A similar calculation establishes

\[
(A + B) \otimes \{1^2\} = A \otimes \{1^2\} + A \otimes B + B \otimes \{1^2\},
\]
and, since both $A \otimes \{1^2\}$ and $B \otimes \{1^2\}$ are one-dimensional $gl_2$ modules and hence also one-dimensional as $sl_2 < gl_2$ and $l_2 < gl_2$ modules, respectively, we find $f_A = 2$.

The evaluation of the explicit quadratic Markov invariants for $L = 2$ proceeds as follows. We regard the probability pattern frequency array $(P_{ij})_{i,j \in \{A, C, G, T\}}$ as an element of $\mathcal{S} \otimes \mathcal{S}$, viz.

$$P = \sum_{i,j} P_{ij} e_i \otimes e_j$$

relative to the standard unit vectors $e_A, e_C, e_G, e_T$ for $\mathcal{S} \cong \mathbb{C}^4$. Via the transformation to the split basis $\{u_0, u_1, v_0', v_1'\}$ we can write (in an obvious notation):

$$\begin{pmatrix}
P_{00} & P_{01} & P_{00'} & P_{01'} \\
P_{10} & P_{11} & P_{10'} & P_{11'} \\
P_{00'} & P_{01'} & P_{10'} & P_{11'} \\
P_{10'} & P_{11'} & P_{10'} & P_{11'}
\end{pmatrix} = \begin{pmatrix}
P_{AA-AT-TA+TT} & P_{AA-AT-TA-TT} & P_{AC-AG-TC+TG} & P_{AC-AG-TC-TG} \\
P_{AA-AT-TA+TT} & P_{AA-AT-TA-TT} & P_{AC-AG+TC+TG} & P_{AC-AG+TC-TG} \\
P_{CA-CT-GA+GT} & P_{CA-CT-GA-GT} & P_{CC-CG-GC+GG} & P_{CC-CG-GC-GG} \\
P_{CA-CT-GA+GT} & P_{CA-CT-GA-GT} & P_{CC-CG+GC+GG} & P_{CC-CG+GC-GG}
\end{pmatrix},$$

where

$$P = \sum P_{ab} u_a \otimes u_b + P_{ab'} u_a \otimes v_{b'} + P_{a'b} v_{a'} \otimes u_b + P_{a'b'} v_{a'} \otimes v_{b'},$$

and the summation is over $a, b \in \{0, 1\}$ and $a', b' \in \{0', 1'\}$.

Using a standard argument, Markov invariants quadratic in the components of the array $P$ can be constructed in the split basis by utilising the antisymmetric Levi-Civita symbols, $\varepsilon_{ab}$ and $\varepsilon_{a'b'}$, where $\varepsilon_{01} = -\varepsilon_{10}$ and $\varepsilon_{00} = \varepsilon_{11} = 0$ and similar for $\varepsilon_{a'b'}$. Clearly there are four possibilities$^8$.

$$I = \sum \varepsilon_{ab} \varepsilon_{cd} P_{ad} P_{bc}, \quad \mathcal{J} = \sum \varepsilon_{ab} \varepsilon_{cd'} P_{ad'} P_{bc'}, \quad \mathcal{T} = \sum \varepsilon_{a'b'} \varepsilon_{cd} P_{ad} P_{b'd'}, \quad \mathcal{F} = \sum \varepsilon_{a'b'} \varepsilon_{c'd'} P_{ad'} P_{b'c'},$$

where each summation is over $a, b, c, d \in \{0, 1\}$ and $a', b', c', d' \in \{0', 1'\}$. We find

$$I = P_{00} P_{11} - P_{10} P_{01}, \quad \mathcal{J} = P_{00'} P_{11'} - P_{10'} P_{01'}, \quad \mathcal{T} = P_{00'} P_{11'} - P_{10'} P_{01'},$$

which can, of course, be recognised as determinants of the four $2 \times 2$ blocks comprising $P$ in the split basis.

Let us evaluate these quadratic Markov invariants on a two-leaf phylogenetic tree. In the general case, parameterise the root distribution as $(p_A, p_C, p_G, p_T) := (p+r, q+s, q-s, p-r)$ with $p+q=1$. Immediately after speciation into two taxa, in the standard basis we have, for each $i, j \in \{A, C, G, T\}$:

$$P_{ij} = \begin{cases} 
p_i, & \text{if } i = j, \\
0, & \text{otherwise};
\end{cases}$$

and hence in the split basis

$$\begin{pmatrix}
P_{00} & P_{01} & P_{00'} & P_{01'} \\
P_{10} & P_{11} & P_{10'} & P_{11'} \\
P_{00'} & P_{01'} & P_{10'} & P_{11'} \\
P_{10'} & P_{11'} & P_{10'} & P_{11'}
\end{pmatrix} = \begin{pmatrix} 2p & 2r & 0 & 0 \\
2r & 2p & 0 & 0 \\
0 & 0 & 2q & 2s \\
0 & 0 & 2s & 2q \end{pmatrix},$$

giving the initial values $I = 4(p^2 - r^2), \mathcal{T} = 4(q^2 - s^2), \mathcal{J} = 0 = \mathcal{F}$.

$^8$The fifth invariant in the count of Lemma 3.1 above is simply the square of the total probability mass.
It is immediately obvious that \( I \) and \( \bar{I} \) are additionally phylogenetic invariants for the model.\(^9\) On the other hand, given the invariance of these functions (the fact that they are one-dimensional representations of the underlying Markov group) means that \( I \) and \( \bar{T} \) can be used to extract distance information from the model, as follows.

Working in the split basis, the strand symmetric Markov matrix \( M \) is cast via Lemma 2 above into the form of a direct sum of two \( 2 \times 2 \) blocks,

\[
M = \begin{pmatrix}
m & 0 \\
0 & m'
\end{pmatrix}.
\]

By a standard argument (the so-called ‘pulley-principle’ \(^6\)), it is enough to evaluate the special case where after speciation one taxon remains fixed whilst the DNA of the other undergoes random substitutions specified generically by \( M \). Correspondingly, the invariants \( I \) and \( \bar{I} \) will undergo rescaling according to

\[
I \to \det(m)I, \quad \bar{T} \to \det(m')\bar{T}.
\]

Recalling Jacobi’s matrix formula: \( \det e^{Qt} = e^{\text{tr}(Q)t} \), and inspection of the diagonal forms given in Lemma 3 shows that the only generators of the Lie algebra with non-zero trace are \( \hat{R} \) and \( \hat{X} \).

Considering the block form and evaluating matrix traces yields

\[
\det(m) = e^{-2(\alpha_3 + \beta_3) + (\alpha_1 + \beta_1 + \alpha_2 + \beta_2)t - \log(p^2 - r^2)},
\]

\[
\det(m') = e^{-(\alpha_1 + \beta_1 + \alpha_2 + \beta_2)t}
\]

Expressing our invariants in the \( \text{LogDet} \) form produces (including the contributions from the initial values just after speciation) finally

\[
- \log\left(\frac{1}{4} I\right) = (2(\alpha_3 + \beta_3) + (\alpha_1 + \beta_1 + \alpha_2 + \beta_2))t - \log(p^2 - r^2),
\]

\[
- \log\left(\frac{1}{4} \bar{T}\right) = (\alpha_1 + \beta_1 + \alpha_2 + \beta_2)t - \log(q^2 - s^2),
\]

which, depending upon one’s willingness to make assumptions about the root distribution parameters \( p, q, r \) and \( s \) (e.g. assuming a stationary distribution), can be used to obtain independent estimators of the sum \( (\alpha_3 + \beta_3) t \), and the sum \( (\alpha_1 + \beta_1 + \alpha_2 + \beta_2) t \).

REFERENCES

[1] E. S. Allman and J. A. Rhodes. Phylogenetic ideals and varieties for the general Markov model. *Adv. Appl. Math.*, to appear, 2007.

[2] D. Barry and J. A. Hartigan. Asynchronous distance between homologous DNA sequences. *Biometrics*, 43:261–276, 1987.

[3] M. Casanellas and S. Sullivant. *Algebraic Statistics for Computational Biology*, chapter The Strand Symmetric Model, pages 305–321. Cambridge University Press, New York, 2005.

[4] Marta Casanellas and Jesús Fernández-Sánchez. Geometry of the Kimura 3-parameter model. *Adv. in Appl. Math.*, 41:265–292, 2008.

[5] Karin Erdmann and Mark J. Wildon. *Introduction to Lie Algebras*. Springer-Verlag, London, 2006.

[6] J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17:368–376, 1981.

[7] M. D. Hendy. The relationship between simple evolutionary tree models and observable sequence data. *Syst. Zool.*, 38:310–321, 1989.

[8] M. D. Hendy, D. Penny, and M. Steel. A discrete Fourier analysis for evolutionary trees. *Proc. Natl. Acad. Sci.*, 91:3339–3343, 1994.

[9] Barbara R. Holland, Jeremy G. Sumner, and Peter D. Jarvis. Low-Parameter Phylogenetic Inference Under the General Markov Model. *Syst. Biol.*, 62:78–92, 2013.

\(^9\)In fact, writing \( S \otimes S = U \otimes U \oplus U \otimes V \oplus V \otimes U \oplus V \otimes V \) we see that the subspaces \( U \otimes V \) and \( V \otimes U \) are actually spanned by linear phylogenetic invariants, with \( J \) and \( \bar{J} \) occurring as quadratic forms constructed from each subspace, respectively.
[10] P. D. Jarvis and J. G. Sumner. Markov invariants for phylogenetic rate matrices derived from embedded submodels. Trans. Comp. Biol. and Bioinf., 9:828–836, 2012.
[11] P. D. Jarvis and J. G. Sumner. Adventures in invariant theory. arXiv:1205.5433 [q-bio.QM], 2013.
[12] J. A. Lake. Reconstructing evolutionary trees from DNA and protein sequences: Paralinear distances. Proceedings of the National Academy of Sciences, 91:1455–1459, 1994.
[13] D. E. Littlewood. The Theory of Group Characters. Clarendon Press, Oxford, 1940.
[14] P. J. Lockhart, M. A. Steel, M. D. Hendy, and D. Penny. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol., 11:605–612, 1994.
[15] Lior Pachter and Bernd Sturmfels, editors. Algebraic Statistics for Computational Biology. Cambridge University Press, New York, 2005.
[16] C. Semple and M. Steel. Phylogenetics. Oxford University Press, 2003.
[17] B. Sturmfels and S. Sullivant. Toric ideals of phylogenetic invariants. J. Comput. Biol., 12:204–228, 2005.
[18] J. G. Sumner, M. A. Charleston, L. S. Jermiin, and P. D. Jarvis. Markov invariants, plethysms, and phylogenetics. J. Theor. Biol., 253:601–615, 2008.
[19] J. G. Sumner, J. Fernández-Sánchez, and P. D. Jarvis. Lie Markov models. J. Theor. Biol., 298:16–31, 2012.
[20] J. G. Sumner, B. R. Holland, and P. D. Jarvis. The algebra of the general Markov model on trees and networks. Bull. Math. Biol., 74(4):858–880, 2012.
[21] J. G. Sumner and P. D. Jarvis. Markov invariants and the isotropy subgroup of a quartet tree. J. Theor. Biol., 258:302–310, 2009.
[22] Jeremy G. Sumner, Jesús Fernández-Sánchez, Peter D. Jarvis, and Michael D. Woodhams. Lie Markov models with purine-pyrimidine symmetry. arXiv:1206.1401 [q-bio.PE], 2013.
[23] Jeremy G. Sumner, Peter D. Jarvis, Jesús Fernández-Sánchez, Bodie T. Kaine, Michael D. Woodhams, and Barbara R. Holland. Is the general time-reversible model bad for molecular phylogenetics? Syst. Biol., 61:1069–74, 2012.
[24] L. A. Székely, M. A. Steel, and P. L. Erdős. Fourier calculus on evolutionary trees. Adv. Appl. Math., 14:200–216, 1993.
[25] S. Tavaré. Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. Lectures on Mathematics in the Life Sciences (American Mathematical Society), 17:57–86, 1986.
[26] H. Weyl. The Theory of Groups and Quantum Mechanics. Dover Publications, 1950.
[27] Von Bing Yap and Lior Pachter. Identification of evolutionary hotspots in the rodent genomes. Genome research, 14(4):574–579, 2004.