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 parameters are unchanged under nucleotide permutations given by Watson-Crick base conjugation). Algebraic analysis of the model's underlying structure as a matrix group leads to a change of basis where the rate generator matrix is given by a two-part block decomposition. We apply representation theoretic techniques and, for any (fixed) number of phylogenetic taxa L and polynomial degree D of interest, provide the means to classify and enumerate the associated Markov invariants. In particular, in the quadratic and cubic cases we prove there are precisely $\frac{1}{3}(3^L+(-1)^L)$ and 6^{L-1} linearly independent Markov invariants, respectively. Additionally, we give the explicit polynomial forms of the Markov invariants for (i) the quadratic case with any number of taxa L, and (ii) the cubic case in the special case of a three-taxa phylogenetic tree. We close by showing our results are of practical interest since the quadratic Markov invariants provide independent estimates of phylogenetic distances based on (i) substitution rates *within* Watson-Crick conjugate pairs, and (ii) 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. In turn, all of these modes of data representation require adequate mathematical model building in stochastic settings in order to capture the essential process systematics with parsimonious parametrizations.

Despite these ongoing challenges, the original brief of phylogenetics – the use of quantitative, inter-species 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. Under further assumptions about absolute mutation rates, parameter estimation then permits recovery of evolutionary divergence times (see [13] for general background on phylogenetic methods).

For nucleic acid base sequence data, the so-called general Markov model is in practice specialized, so that the key theoretical object – an assumed 4×4 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 [35]; further constraints on the parameters lead to one of a number of other model types. Amongst these, we distinguish the "group-based" models ([27],

Date: July 2013.

2000 *Mathematics Subject Classification.* Primary 16W30; Secondary 05E05;

chapter 8), which allow for direct analytical treatments, using discrete Fourier or Hadamard inversion techniques [15, 34].

The armoury of theoretical techniques has been further enriched with the advent of algebraically-inspired methods which seek to locate certain geometric structures, 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 the relatively new field of "algebraic statistics" [26].

Turning to computational approaches, although maximum likelihood optimization is powerful enough to allow full parameter recovery, in principle even for the general Markov model [7], in practical implementations it is usual to work with specialized models. In [30] we argued for the natural criterion of *closure* (under matrix multiplication) as a guide to model choice in phylogenetics. In that work 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 [33]. 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 [30]. In the continuous time context, these models have affiliated Lie algebras where the rate matrices are contained within an appropriate stochastic cone (see [14] for details).

Of course, the general Markov model itself is by construction multiplicatively closed, and in related work [29, 32] 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 [32, 17].

Our work on Markov invariants must be distinguished from related work on the similarly named *phylogenetic invariants* [21, 6, 12, 10]. Phylogenetic invariants are defined as those polynomials that vanish on a given phylogenetic tree (or subset of trees) under all (or nearly all) parameter settings of a given Markov model of sequence evolution. As such, phylogenetic invariants form polynomial *ideals* and hence can be analysed formally using algebraic geometry [1, 28, 8, 5]. Beyond the theoretical significance of phylogenetic invariants (for example, they can be used to establish model identifiability [2]), the practical motivation behind the development of phylogenetic invariants lies in their vanishing (at least in expectation value) on particular trees. Thus, when evaluated on an observed sequence alignment, phylogenetic invariants provide some information as to which evolutionary tree history the sequences are likely to have arisen from.

On the other hand, Markov invariants are defined as the one-dimensional polynomial *representations* of the matrix group formed from the Markov matrices that act on the leaves of a phylogenetic tree. By definition, each Markov invariant spans a one-dimensional invariant subspace under changes of model parameter settings at the leaves of the tree. Hence, Markov invariants provide useful statistical information that is invariant to the independent stochastic processes that have occurred since phylogenetically related taxa diverged from one another. Phylogenetic invariants do not share this invariance property, and it is our contention that, at least comparatively, Markov invariants will provide particularly robust statistical information (particularly if we consider the setting of finite length sequence alignments where stochastic errors become important).

In a study of rodent phylogeny, a hitherto un-noticed interesting regime of DNA substitution parameters was pointed out by Yap and Pachter [38]. 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 [4, 5].

This article focusses exclusively on a representation theoretic approach to the strand symmetric model and the derivation of Markov invariants for this model. This is achieved by exploring a formal algebraic analysis of the Lie algebra associated with the model. In §2, we provide an abstract decomposition of this Lie algebra in terms of the Lie algebras of classical groups [36], and identify the particular representation provided by the 4×4 rate matrices making up strand symmetric model. In §3, we couple our previous work characterising Markov invariants for the general Markov model [29], and our analysis of the underlying Lie algebra in §2, to provide a complete classification and enumeration of binary and cubic Markov invariants for the strand symmetric model. The most technical aspects of the classification and enumeration of Markov invariants – relying heavily on specialised manipulations of symmetric function characters (plethysm and skew operations) – are relegated to the appendix §A; the casual reader should be able to follow the explicit construction of the invariants, without the need to fully understand the combinatorial derivations underlying our enumerations. In §4, we examine the evaluation of quadratic Markov invariants on for a two-leaf phylogenetic tree. In this case there are four quadratic Markov invariants, which we show provide the means for estimating *two* pairwise phylogenetic distances: constructed from the total of substitution rates *within*, and *across*, Watson-Crick conjugate base pairs, respectively. In the discussion §5, we give concluding remarks and possibilities for future work including a comparison of the relative statistical power of phylogenetic and Markov invariants to accurately recover evolutionary trees.

Acknowledgement

Part of this work was completed by PDJ 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 $\{\mathsf{R},\mathsf{Y}\}$ (purines and pyrimidines), or a 4 state system with state space $\{\mathsf{A},\mathsf{C},\mathsf{G},\mathsf{T}\}$.

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

$$M=\left(\begin{array}{cc} m_{\mathsf{RR}} & m_{\mathsf{RY}} \\ m_{\mathsf{YR}} & m_{\mathsf{YY}}\end{array}\right).$$

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

$$M(a,b)=\left(\begin{array}{cc} 1-a & a \\ b & 1-b\end{array}\right).$$

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 $0\leq a,b\leq 1$ (technically a matrix semigroup). 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 2×2 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[1]

$$R_1 = \begin{pmatrix} \overline{1} & 1 \\ 0 & 0 \end{pmatrix}, \qquad R_2 = \begin{pmatrix} 0 & 0 \\ 1 & \overline{1} \end{pmatrix},$$

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 $\{\mathsf{A,C,G,T}\}$, is similar. Allowing for the row sum constraint, the Markov matrix

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

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 of 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, with the general Markov rate model as above. We refer to these rate matrices as the stochastic cone of the Lie algebra[2].

Consider now the general time reversible (GTR) model, where the guiding assumption is that transition rates involving arbitrary states $i, j \in \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$, weighted by the (stationary) distribution of the starting state π_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,

$$Q = \begin{pmatrix} Q_{\mathsf{AA}} & S_{\mathsf{AC}}\pi_{\mathsf{C}} & S_{\mathsf{AG}}\pi_{\mathsf{G}} & S_{\mathsf{AT}}\pi_{\mathsf{T}} \\ S_{\mathsf{CA}}\pi_{\mathsf{A}} & Q_{\mathsf{CC}} & S_{\mathsf{CG}}\pi_{\mathsf{G}} & S_{\mathsf{CT}}\pi_{\mathsf{T}} \\ S_{\mathsf{GA}}\pi_{\mathsf{A}} & S_{\mathsf{GC}}\pi_{\mathsf{C}} & Q_{\mathsf{GG}} & S_{\mathsf{GT}}\pi_{\mathsf{T}} \\ S_{\mathsf{TA}}\pi_{\mathsf{A}} & S_{\mathsf{TC}}\pi_{\mathsf{C}} & S_{\mathsf{TG}}\pi_{\mathsf{G}} & Q_{\mathsf{TT}} \end{pmatrix},$$

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

[1] For aesthetic purposes here and below, signed entries in matrices are written with overbars.

[2] Consult [14]; details of the general case are not required in the present work.

A special case of the GTR model occurs when its transition rates are unchanged under Watson-Crick base pairing conjugation (i.e. $\mathsf{A} \leftrightarrow \mathsf{T}$, $\mathsf{C} \leftrightarrow \mathsf{G}$); for example $Q_{\mathsf{CA}} = Q_{\mathsf{GT}}$, $Q_{\mathsf{CT}} = Q_{\mathsf{GA}}$, $Q_{\mathsf{TA}} = Q_{\mathsf{AT}}$, and so on. In the above parametrization, imposition of this constraint on self-conjugate pairs such as $Q_{\mathsf{TA}} = Q_{\mathsf{AT}}$ enforces Chargaff's rule on the stationary distribution, $\pi_{\mathsf{A}} = \pi_{\mathsf{T}}$, and $\pi_{\mathsf{C}} = \pi_{\mathsf{G}}$, and the remaining conditions constrain S also to fulfil the analogous conditions $S_{\mathsf{AC}} = S_{\mathsf{GT}}$, $S_{\mathsf{AG}} = S_{\mathsf{CT}}$ etc. (for self-conjugate pairs, the relations $S_{\mathsf{CG}} = S_{\mathsf{GC}}$ and $S_{\mathsf{AT}} = S_{\mathsf{TA}}$ are already enforced by the symmetry of S). As mentioned, the GTR model class is not multiplicatively closed [30], 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 alone, 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

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

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' [8], which is are useful generalisation of the standard 'group-based' models ([27], 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 $\{\epsilon, (\mathsf{AT})(\mathsf{GC})\}$ (where ϵ is 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 [30, 14] 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 only 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 [30], 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 [14] a complete hierarchy consisting of 35 multiplicatively closed models is derived[3], which are additionally invariant under the permutations which fix the partitioning of nucleotides into purines and pyrimidines, i.e. $\mathsf{AG}|\mathsf{CT} := \{\{\mathsf{A},\mathsf{G}\},\{\mathsf{C},\mathsf{T}\}\}$, so notationally $\mathsf{AG}|\mathsf{CT} \equiv \mathsf{GA}|\mathsf{CT} \equiv \mathsf{TC}|\mathsf{AG}\ldots$ etc. In [14] it is also noted that an equivalent hierarchy exists for the partitioning that defines the Watson-Crick base pairing conjugation, i.e. $\mathsf{AT}|\mathsf{GC}$ (and yet another hierarchy for the partitioning $\mathsf{AC}|\mathsf{GT}$). In particular, Model 6.6 [14] is identical to the strand symmetric model with the substitution

[3] The 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.ma1.upc.edu/ jfernandez/LMNR.pdf`.

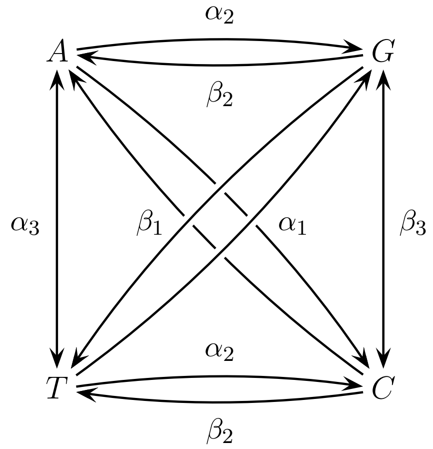

FIGURE 1. Graphical representation of the strand symmetric model.

$\mathsf{G} \leftrightarrow \mathsf{T}$ (or $\mathsf{A} \leftrightarrow \mathsf{C}$). From this point of view, the strand symmetric model lives in a large hierarchy of Lie Markov models, equivalent to the hierarchy presented in [14], where each model has symmetry consistent with Watson-Crick base pairing.

Let $\mathcal{S}$ be the vector space associated with the four nucleotide bases, with standard unit vectors

$$e_{\mathsf{A}} = (1,0,0,0)\,, \quad e_{\mathsf{C}} = (0,1,0,0)\,, \quad e_{\mathsf{G}} = (0,0,1,0)\,, \quad e_{\mathsf{T}} = (0,0,0,1)\,,$$

so $\mathcal{S} := \langle e_{\mathsf{A}}, e_{\mathsf{C}}, e_{\mathsf{G}}, e_{\mathsf{T}} \rangle_{\mathbb{C}} \cong \mathbb{C}^4$ and, for example, the state distribution π_i is given by the vector $\pi = \pi_{\mathsf{A}} e_{\mathsf{A}} + \pi_{\mathsf{C}} e_{\mathsf{C}} + \pi_{\mathsf{G}} e_{\mathsf{G}} + \pi_{\mathsf{T}} e_{\mathsf{T}}$. 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:

$$S_1 = \begin{pmatrix} \bar{1} & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & \bar{1} \end{pmatrix}, \qquad S_2 = \begin{pmatrix} \bar{1} & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & \bar{1} \end{pmatrix}, \qquad S_3 = \begin{pmatrix} \bar{1} & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & \bar{1} \end{pmatrix},$$
$$T_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & \bar{1} & 0 & 0 \\ 0 & 0 & \bar{1} & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad T_2 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \bar{1} & 0 & 1 \\ 1 & 0 & \bar{1} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad T_3 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \bar{1} & 1 & 0 \\ 0 & 1 & \bar{1} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

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 *Ansatz* $\pi_{\mathsf{A}} = \pi_{\mathsf{T}} = p$, $\pi_{\mathsf{C}} = \pi_{\mathsf{G}} = q$ 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 α_3 and β_3 – which is therefore the unique stationary distribution. A graphical representation of the model is given in Figure 1.

The full set of 15 commutation relations amongst these generators is

$$\begin{aligned}
&[S_1,S_2] = S_1 - S_2, \qquad [S_2,S_3] = -S_1 + S_2, \qquad [S_3,S_1] = -S_1 + S_2,\\
&[T_1,T_2] = T_1 - T_2, \qquad [T_2,T_3] = -T_1 + T_2, \qquad [T_3,T_1] = -T_1 + T_2,\\
&[S_1,T_1] = -S_1 + T_1, \quad [S_1,T_2] = -S_1 + S_3 - T_3 + T_2, \quad [S_1,T_3] = -S_1 + S_2,\\
&[S_2,T_1] = -S_2 + S_3 + T_1 - T_3, \quad [S_2,T_2] = -S_2 + T_2, \quad [S_2,T_3] = S_1 - S_2,\\
&[S_3,T_1] = T_1 - T_2, \quad [S_3,T_2] = -T_1 + T_2, \qquad [S_3,T_3] = 0,
\end{aligned}$$

as can be checked by elementary matrix algebra. We denote the corresponding complex Lie algebra by $l_{\mathsf{SSM}} := \langle S_1, S_2, S_3, T_1, T_2, T_3 \rangle_{\mathbb{C}}$.

The group of permutations $\{\epsilon, (\mathsf{AT}), (\mathsf{GC}), (\mathsf{AT})(\mathsf{GC}), (\mathsf{AG})(\mathsf{CT}), (\mathsf{AC})(\mathsf{GT}), (\mathsf{AGTC}), (\mathsf{ATGC})\}$ fix the Watson-Crick pairing $\mathsf{AT}|\mathsf{GC}$, and are generated, for example, by the permutation (AT), via $\mathsf{TA}|\mathsf{GC} \equiv \mathsf{AT}|\mathsf{GC}$, and the permutation $(\mathsf{AG})(\mathsf{TC})$, via $\mathsf{GC}|\mathsf{AT} \equiv \mathsf{AT}|\mathsf{GC}$. 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 [9] 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 a vector space) as gl_1. As generators of the so-called 'defining' representation of sl_2, with corresponding module $\mathcal{U} \cong \mathbb{C}^2$, we take:

$$K_+ = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \qquad K_- = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \qquad K_0 = \tfrac{1}{2}\begin{pmatrix} 1 & 0 \\ 0 & \overline{1} \end{pmatrix},$$

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 [9], 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[4] 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.

$$X = \tfrac{1}{2}\begin{pmatrix} \overline{1} & 1 \\ 1 & \overline{1} \end{pmatrix}, \quad Y = \tfrac{1}{2}\begin{pmatrix} 1 & \overline{1} \\ 1 & \overline{1} \end{pmatrix}.$$

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 $\mathcal{R} \cong \mathbb{C} \cong \langle v \rangle_{\mathbb{C}}$ and v is an eigenvector of a generator R of gl_1 with eigenvalue r. 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 corresponding module $\mathcal{V}' = \langle v' \rangle_{\mathbb{C}} \cong \mathbb{C}$, where v' is an eigenvector for X and is annihilated by Y.

Lemma 1: **Decomposition of the Lie algebra of the strand symmetric model**
The Lie algebra l_{SSM} generated by $S_1, S_2, S_3, T_1, T_1, T_2$ 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.

[4] See [31] for an algebraic investigation of the role of l_2 in phylogenetic tree and network models.

Proof: Define the new set of generators,

$$
\begin{aligned}
&\widehat{K}_0 = \tfrac{1}{4}(-S_3+T_3), \quad \widehat{K}_+ = \tfrac{1}{2}(S_1-S_2), \quad \widehat{K}_- = \tfrac{1}{2}(T_1-T_2);\\
&\widehat{R} = \tfrac{1}{2}(S_3+T_3);\\
&\widehat{X} = \tfrac{1}{4}(S_1+S_2+T_1+T_2), \qquad \widehat{Y} = \tfrac{1}{4}(-S_1-S_2+S_3+T_1-T_3+T_2).
\end{aligned}
$$

By direct computation,

$$
\widehat{K}_0 = \tfrac{1}{4}\begin{pmatrix} 1 & 0 & 0 & \bar{1}\\ 0 & \bar{1} & 1 & 0\\ 0 & 1 & \bar{1} & 0\\ \bar{1} & 0 & 0 & 1 \end{pmatrix}, \quad
\widehat{K}_+ = \tfrac{1}{2}\begin{pmatrix} 0 & 1 & \bar{1} & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & \bar{1} & 1 & 0 \end{pmatrix}, \quad
\widehat{K}_- = \tfrac{1}{2}\begin{pmatrix} 0 & 0 & 0 & 0\\ 1 & 0 & 0 & \bar{1}\\ \bar{1} & 0 & 0 & 1\\ 0 & 0 & 0 & 0 \end{pmatrix};
$$

$$
\widehat{R} = \tfrac{1}{2}\begin{pmatrix} \bar{1} & 0 & 0 & 1\\ 0 & \bar{1} & 1 & 0\\ 0 & 1 & \bar{1} & 0\\ 1 & 0 & 0 & \bar{1} \end{pmatrix}; \quad
\widehat{X} = \tfrac{1}{4}\begin{pmatrix} \bar{2} & 1 & 1 & 0\\ 1 & \bar{2} & 0 & 1\\ 1 & 0 & \bar{2} & 1\\ 0 & 1 & 1 & \bar{2} \end{pmatrix}, \quad
\widehat{Y} = \tfrac{1}{4}\begin{pmatrix} 1 & \bar{1} & \bar{1} & 1\\ 1 & \bar{1} & \bar{1} & 1\\ 1 & \bar{1} & \bar{1} & 1\\ 1 & \bar{1} & \bar{1} & 1 \end{pmatrix},
$$

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

□

Lemma 2: **Decomposition of the state space $\mathcal{S}$ of strand symmetric model**
As a module of $l_{\mathsf{SSM}} \cong sl_2 \oplus gl_1 \oplus l_2$, the four-dimensional state space $\mathcal{S}$ decomposes as the direct sum of two two-dimensional components $\mathcal{S} = U \oplus V$ where

$$
\begin{aligned}
U &\cong \left(\mathcal{U} \otimes \mathcal{R} \otimes \mathcal{V}'\right),\\
V &\cong \left(\mathcal{U}_0 \otimes \mathcal{R}_0 \otimes \mathcal{V}\right),
\end{aligned}
$$

and

(1) $\mathcal{U} \cong \mathbb{C}^2$ and $\mathcal{U}_0 \cong \mathbb{C}^1$ are the sl_2 modules described above,
(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 $\{u_0, u_1, v_{0'}, v_{1'}\}$[5] where

$$
u_0 = \left(1,0,0,\overline{1}\right), \quad u_1 = \left(0,1,\overline{1},0\right), \quad v_{0'} = (1,0,0,1), \quad v_{1'} = (0,1,1,0).
$$

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 2×2 identity matrix):

$$
\widehat{K}_0 = \begin{pmatrix} K_0 & 0\\ 0 & 0 \end{pmatrix}, \quad \widehat{K}_+ = \begin{pmatrix} K_+ & 0\\ 0 & 0 \end{pmatrix}, \quad \widehat{K}_- = \begin{pmatrix} K_- & 0\\ 0 & 0 \end{pmatrix},
$$

$$
\widehat{R} = \begin{pmatrix} -\mathbf{1} & 0\\ 0 & 0 \end{pmatrix}, \quad \widehat{X} = \begin{pmatrix} -\frac{1}{2}\mathbf{1} & 0\\ 0 & X \end{pmatrix}, \qquad \widehat{Y} = \begin{pmatrix} 0 & 0\\ 0 & Y \end{pmatrix}.
$$

[5] Here and below we will mark the indices of vectors (and/or tensor components) in V by $'$.

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. Application to Markov invariants

Our aim thus far has been to present the strand symmetric model [38, 4] from the point of view of the underlying continuous Lie group. In Lemma 1 we gave a classical decomposition 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 applications of these results.

In the analysis of [4], the self-similar structure of the substitution matrices 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 Fourier 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` [3, 22, 24]. The `LogDet` arises as a particular example of the more general concept of *Markov invariants* [29], 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 [16], 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_{SSM}), a complete set of linear invariants is not available (and hence no linear inversion technique analagous to the Hadamard conjugation is applicable); 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, §A, which relies on our previously established rules for working out the appropriate representations [19] using calculations in the ring of symmetric functions. Here we quote the main result:

Theorem 1: **Count of quadratic Markov invariants for the strand symmetric model**
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[6].

Proof: See §A below. □

In the $L = 2$ leaf case the quadratic invariants are proxies for determinant functions, not of the full 4×4 probability array, but for its 2×2 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

[6] Integer sequence A046717 (see `http://oeis.org/`).

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 below in §4 and should be compared to application of the `LogDet` [3, 22, 24], which, when considered as a distance measure for the strand symmetric model, conjoins these two quantities into a total sum.

In general, the $\frac{1}{2}(3^L+(-1)^L)$ quadratic Markov invariants can be constructed as follows. We work in the split basis $\{u_0,u_1,v_{0'},v_{1'}\}$ and define the anti-symmetric tensors

$$\epsilon_{ij}:=\begin{cases}\;\;1 \text{ if } i=0, j=1,\\ -1 \text{ if } i=1, j=0,\\ \;\;0, \qquad \text{otherwise;}\end{cases}\qquad \overline{\epsilon}_{ij}:=\begin{cases}\;\;1 \text{ if } i=0', j=1',\\ -1 \text{ if } i=1', j=0',\\ \;\;0, \qquad \text{otherwise.}\end{cases}$$

Given the block form $M=m\oplus\overline{m}$ of a strand symmetric model substitution matrix in the split basis, we have

$$\begin{aligned}\det(m)&=\sum_{i_1,i_2,j_1,j_2=0,1}M_{i_1i_2}M_{j_1j_2}\epsilon_{i_1j_1}\epsilon_{i_2j_2},\\ \det(\overline{m})&=\sum_{i_1,i_2,j_1,j_2=0',1'}M_{i_1i_2}M_{j_1j_2}\overline{\epsilon}_{i_1j_1}\overline{\epsilon}_{i_2j_2}.\end{aligned}$$

Choose an integer $q\leq L$ and a sequence $(a_1,a_2,\ldots,a_q)$ with $a_j\in\{1,2\}$, and consider the quadratic function $f^{(a_1,a_2,\ldots a_q)}$ on L-way tensors $\psi_{i_1i_2i_3\ldots i_L}$ defined by:

$$f^{(a_1,a_2,\ldots a_q)}(\psi):=\sum\psi_{i_1i_2i_3\ldots i_q0'0'\ldots0'}\psi_{j_1j_2j_3\ldots j_q0'0'\ldots0'}\epsilon^{(a_1)}_{i_1j_1}\epsilon^{(a_2)}_{i_2j_2}\ldots\epsilon^{(a_q)}_{i_qj_q},$$

where $\epsilon^{(1)}_{ij}\equiv\epsilon_{ij}$ and $\epsilon^{(2)}_{ij}\equiv\overline{\epsilon}_{ij}$ and the summation is over the values $0,1,0',1'$ for all indices appearing in the expression. An explicit check shows that if $\psi\rightarrow\psi'=M_1\otimes M_2\otimes\ldots M_L\cdot\psi$, where each M_i a strand symmetric model substitution matrix, we have

$$f^{(a_1,a_2,\ldots,a_q)}(\psi')=\widehat{\zeta}f^{(a_1,a_2,\ldots,a_q)}(\psi),$$

with $\widehat{\zeta}:=\prod_{1\leq i\leq q}\det(m^{(a_i)}_i)$ and $m^{(1)}_i\equiv m_i$ and $m^{(2)}_i\equiv\overline{m}_i$. From the multiplicative property of the determinant, it follows that each such function provides a Markov invariant for the strand symmetric model. Allowing for analogous constructions utilizing different subsets of q parts of the tensor $\psi_{i_1i_2\ldots i_L}$ (and marginalizing on the remaining $L-q$ parts), shows that we can construct

$$\sum_{q=0}^{L}\binom{L}{q}2^q,$$

Markov invariants in this way. However, it is easy to show, using the anti-symmetry of ϵ_{ij} and $\overline{\epsilon}_{ij}$, that if q is odd the construction gives the zero polynomial. For example, for $L=3$, we have

$$\begin{aligned}f^{(1,1,2)}(\psi)&=\psi_{000'}\psi_{111'}-\psi_{100'}\psi_{011'}+\psi_{110'}\psi_{001'}-\psi_{010'}\psi_{101'}-\psi_{001'}\psi_{110'}\\ &\qquad+\psi_{101'}\psi_{010'}-\psi_{111'}\psi_{000'}+\psi_{011'}\psi_{100'}\\ &=0.\end{aligned}$$

Excluding the cases where q is odd, we see that we have constructed

$$\sum_{q=0}^{L}\binom{L}{2q}2^{2q}=\frac{1}{3}(3^{L}+(-1)^{L}),$$

Markov invariants that clearly linear independent (since they have distinct weights $\widehat{\zeta}$) consistent with Theorem 1, as required.

The reader should note that the construction of the binary Markov invariants can easily be understood in intuitive terms by taking an $L-$way tensor $\psi_{i_1i_2i_3\ldots i_L}$ and implementing two steps, as follows. Firstly, we "marginalize" $(L-q)$ of the indices via,

$$\psi_{i_1i_2i_3\ldots i_L}\rightarrow\psi_{i_1i_2i_3\ldots i_q0'0'\ldots 0'},$$

where the $0'$ component simply expresses probability conservation in the underlying Markov chain. Secondly, we exploit the block form $M=m\oplus\overline{m}$ of the strand symmetric model by considering $m,\overline{m}\in GL(2)$ and "saturate" indices with the $GL(2)$ invariant tensors ϵ_{ij} and $\overline{\epsilon}_{ij}$. As such, beyond the initial marginalization step, there is no direct exploitation of the more fined-grained observation that $\overline{m}$ actually belongs to a proper matrix-subgroup of $GL(2)$.

However, in the higher degree Markov invariants, things become combinatorially more interesting, as we now illustrate specifically for the cubic case.

Theorem 2: **Count of cubic Markov invariants for the strand symmetric model**
For L leaves there are precisely 6^{L-1} linearly independent cubic Markov invariants for the strand symmetric model, namely $1,6,36,216,\ldots$ for $L=1,2,3,4,\ldots$ respectively.

Proof:
See § A below. □

As alluded to above, the explicit construction of the cubic invariants is not as straightforward as the the quadratic case. To illustrate, we give the complete list of 36 cubic linear independent Markov invariants for $L=3$.

Consider the three-way tensors $\psi_{i_1i_2i_3}$ with $\psi\mapsto M_1\otimes M_2\otimes M_3\cdot\psi$ where each $M_i=m_i\oplus\overline{m}_i$ belongs to the strand symmetric model. We set $w_i:=\det(m_i)$ and $\lambda_i:=\det(\overline{m}_i)$. The enumeration given in §A shows that the individual counts for various weights $\widehat{\zeta}$ of the 36 Markov invariants are given by

$$1;\lambda_i\lambda_j;\lambda_1\lambda_2\lambda_3;\lambda_iw_j;2\lambda_i\lambda_jw_k;2w_iw_j;3\lambda_iw_jw_k;4w_1w_2w_3,$$

for all choices $\{i,j,k\}=\{1,2,3\}$ and multiplicities have been included as a multiplicative factor, e.g. there is *one* Markov invariant with weight $\lambda_1\lambda_2$ and there are *three* Markov invariants with weight $\lambda_1w_2w_3$.

Using constructions inspired by the quadratic case above, we used Mathematica [37] to explicitly find 36 linearly independent Markov invariants, with the following results:

(1) The **single** invariant with trivial weight 1 is simply given by the cube of the probability sum: $\psi_{0'0'0'}^{3}$.
(2) For the three quadratic weights of the form $\lambda_i\lambda_j$, the expression $\psi_{0'0'0'}\times\sum\psi_{i_1i_20'}\psi_{j_1j_20'}\overline{\epsilon}_{i_1j_1}\overline{\epsilon}_{i_2j_2}$ plus the obvious two permutations across the tensor indices provides the required **three** invariants.

(3) For the quadratic weight $\lambda_1 w_2$, the expression $\psi_{0'0'0'} \times \sum \psi_{i_1 i_2 0'} \psi_{j_1 j_2 0'} \bar{\epsilon}_{i_1 j_1} \epsilon_{i_2 j_2}$ provides the required invariant. Since there are six distinct quadratic weights of the form $\lambda_i w_j$, this gives the required total of **six** invariants.

(4) For the quadratic weight $w_1 w_2$, the expressions $\psi_{0'0'0'} \times \sum \psi_{0' i_2 i_3} \psi_{0' j_2 j_3} \epsilon_{i_2 j_2} \epsilon_{i_3 j_3}$ and $\sum \psi_{0'0' i_3} \psi_{0' j_2 0'} \psi_{0' k_2 k_3} \epsilon_{j_2 k_2} \epsilon_{i_3 k_3}$ give two linearly independent invariants. Since there are three distinct quadratic weights of the form $w_i w_j$, this gives the required total of **six** invariants.

(5) For the cubic weight $\lambda_1 \lambda_2 w_3$, the expression $\psi_{i_1 i_2 i_3} \psi_{j_1 j_2 0'} \psi_{0'0' k_3} \bar{\epsilon}_{i_1 j_1} \bar{\epsilon}_{i_2 j_2} \epsilon_{i_3 k_3}$ plus two permutations provides only two linearly independent invariants. Since there are three distinct cubic weights of the form $\lambda_i \lambda_j w_k$, this gives the required total of **six** invariants.

(6) For the cubic weight $\lambda_1 w_2 w_3$, the expressions $\sum \psi_{i_1 i_2 i_3} \psi_{j_1 j_2 0'} \psi_{0'0' k_3} \bar{\epsilon}_{i_1 j_1} \epsilon_{i_2 j_2} \epsilon_{i_3 k_3}$, $\sum \psi_{i_1 i_2 i_3} \psi_{j_1 0' j_3} \psi_{0' k_2 0'} \bar{\epsilon}_{i_1 j_1} \epsilon_{i_2 k_2} \epsilon_{i_3 j_3}$ and $\sum \psi_{i_1 i_2 i_3} \psi_{0' j_2 j_3} \psi_{k_1 0' 0'} \bar{\epsilon}_{i_1 k_1} \epsilon_{i_2 j_2} \epsilon_{i_3 j_3}$ provide three linearly independent invariants. Since there are three distinct cubic weights of the form $\lambda_i w_j w_k$, this gives the required total of **nine** invariants.

(7) For the cubic weight $\lambda_1 \lambda_2 \lambda_3$, the expression $\sum \psi_{i_1 i_2 i_3} \psi_{j_1 j_2 0'} \psi_{0'0' k_3} \bar{\epsilon}_{i_1 j_1} \bar{\epsilon}_{i_2 j_2} \bar{\epsilon}_{i_3 k_3}$ plus permutations across tensor indices provides only **one** linearly independent invariant, as required.

(8) For the cubic weight $w_1 w_2 w_3$, the invariant $\sum \psi_{i_1 i_2 i_3} \psi_{j_1 j_2 0'} \psi_{0'0' k_3} \epsilon_{i_1 j_1} \epsilon_{i_2 j_2} \epsilon_{i_3 j_3}$ plus two permutations across tensor indices, plus the expression $\sum \psi_{0' i_2 i_3} \psi_{j_1 0' j_3} \psi_{k_1 k_2 0'} \epsilon_{j_1 k_1} \epsilon_{i_2 k_2} \epsilon_{i_3 j_3}$ provide the required **four** invariants.

The enumeration of the 36 Markov invariants is summarised in Table 3. The reader should note that the first invariant is simply the trivial invariant cubed and the next three lines of invariants are all of the form (trival) × (quadratic). The remaining invariants are non-factorizable, and illustrate how the probability conservation invariance for different parts of the tensor is shared across different terms in the cubic product.

As an example, consider the single invariant with weight $\lambda_1 \lambda_2 \lambda_3$:

$$\sum \psi_{i_1 i_2 i_3} \psi_{j_1 0' j_3} \psi_{0' k_2 0'} \bar{\epsilon}_{i_1 j_1} \bar{\epsilon}_{i_2 k_2} \bar{\epsilon}_{i_3 j_3} = -2 \psi_{0'0'1'} \psi_{1'0'0'} \psi_{0'1'0'} - \psi^2_{0'0'0'} \psi_{1'1'1'} \\ + \psi_{0'0'0'} \left(\psi_{0'1'1'} \psi_{1'0'0'} + \psi_{1'0'1'} \psi_{0'0'1'} + \psi_{1'1'0'} \psi_{0'0'1'} \right).$$

This invariant is of particular interest as it is already known in the context of the two-state general Markov model as both (i) a Markov invariant the "stangle" [29], and (ii) a three-way covariance on triplet trees [20]. The fact that this Markov invariant arises again in the case of the strand symmetric is remarkable, but should be expected given the two-part block decomposition given in Lemma 2.

4. Evaluation of the quadratic Markov invariants on phylogenetic trees

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

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

relative to the standard unit vectors $e_{\mathsf{A}}, e_{\mathsf{C}}, e_{\mathsf{G}}, e_{\mathsf{T}}$ 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

$$\begin{pmatrix} P_{00} & P_{01} & P_{00'} & P_{01'} \\ P_{10} & P_{11} & P_{10'} & P_{11'} \\ P_{0'0} & P_{0'1} & P_{0'0'} & P_{0'1'} \\ P_{1'0} & P_{1'1} & P_{1'0'} & P_{1'1'} \end{pmatrix} = \begin{pmatrix} P_{\mathsf{AA-AT-TA+TT}}, & P_{\mathsf{AC-AG-TC+TG}}, & P_{\mathsf{AA+AT-TA-TT}}, & P_{\mathsf{AC+AG-TC-TG}} \\ P_{\mathsf{CA-CT-GA+GT}}, & P_{\mathsf{CC-CG-GC+GG}}, & P_{\mathsf{CA+CT-GA-GT}}, & P_{\mathsf{CC+CG-GC-GG}} \\ P_{\mathsf{AA-AT+TA-TT}}, & P_{\mathsf{AC-AG+TC-TG}}, & P_{\mathsf{AA+AT+TA+TT}}, & P_{\mathsf{AC+AG+TC+TG}} \\ P_{\mathsf{CA-CT+GA-GT}}, & P_{\mathsf{CC-CG+GC-GG}}, & P_{\mathsf{CA+CT+GA+GT}}, & P_{\mathsf{CC+CG+GC+GG}} \end{pmatrix},$$

where, for instance, $P_{\mathsf{AA+AT-TA-TT}} := P_{\mathsf{AA}} + P_{\mathsf{AT}} - P_{\mathsf{TA}} - P_{\mathsf{TT}}$.

Example Invariant	Weight	Perms	Lin. indep.	Weight perms	Total
$\psi^3_{0'0'0'}$	1	1	1	1	1
$\psi_{0'0'0'}\times\sum\psi_{i_1i_20'}\psi_{j_1j_20'}\bar{\epsilon}_{i_1j_1}\bar{\epsilon}_{i_2j_2}$	$\lambda_1\lambda_2$	1	1	3	3
$\psi_{0'0'0'}\times\sum\psi_{i_1i_20'}\psi_{j_1j_20'}\bar{\epsilon}_{i_1j_1}\epsilon_{i_2j_2}$	λ_1w_2	1	1	6	6
$\psi_{0'0'0'}\times\sum\psi_{0'i_2i_3}\psi_{0'j_2j_3}\epsilon_{i_2j_2}\epsilon_{i_3j_3}$	w_1w_2	1	1	3	3
$\sum\psi_{0'0'i_3}\psi_{0'j_20'}\psi_{0'k_2k_3}\epsilon_{j_2k_2}\epsilon_{i_3k_3}$	w_1w_2	1	1	3	3
$\sum\psi_{i_1i_2i_3}\psi_{j_10'j_3}\psi_{0'k_20'}\bar{\epsilon}_{i_1j_1}\bar{\epsilon}_{i_2k_2}\epsilon_{i_3j_3}$	$\lambda_1\lambda_2w_3$	3	2	3	6
$\sum\psi_{i_1i_2i_3}\psi_{j_10'j_3}\psi_{0'k_20'}\epsilon_{i_1j_1}\epsilon_{i_2k_2}\bar{\epsilon}_{i_3j_3}$	$\lambda_1w_2w_3$	3	3	3	9
$\sum\psi_{i_1i_2i_3}\psi_{j_10'j_3}\psi_{0'k_20'}\epsilon_{i_1j_1}\epsilon_{i_2k_2}\epsilon_{i_3j_3}$	$w_1w_2w_3$	3	3	3	3
$\sum\psi_{0'i_2i_3}\psi_{j_10'j_3}\psi_{k_1k_20'}\epsilon_{j_1k_1}\epsilon_{i_2k_2}\epsilon_{i_3j_3}$	$w_1w_2w_3$	1	1	1	1
$\sum\psi_{i_1i_2i_3}\psi_{j_10'j_3}\psi_{0'k_20'}\bar{\epsilon}_{i_1j_1}\bar{\epsilon}_{i_2k_2}\bar{\epsilon}_{i_3j_3}$	$\lambda_1\lambda_2\lambda_3$	3	1	1	1
					36

TABLE 1. Cubic Markov invariants for 3-way tensors under the strand symmetric model.

Using the notation from above, the five quadratic Markov invariants for $L=2$ are given by (setting $q=0$) the trivial invariant $P^2_{0'0'}=(P_{\mathsf{AA}}+P_{\mathsf{AC}}+\ldots+P_{\mathsf{TT}})^2$ and (setting $q=2$) the four invariants

$$\begin{aligned} f^{(1,1)}(P)&=P_{00}P_{11}-P_{01}P_{10}, & f^{(1,2)}(P)&=P_{00'}P_{11'}-P_{01'}P_{10'},\\ f^{(2,1)}(P)&=P_{0'0}P_{1'1}-P_{0'1}P_{1'0}, & f^{(2,2)}(P)&=P_{0'0'}P_{1'1'}-P_{0'1'}P_{1'0'}, \end{aligned}$$

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

Letting $P\to P'=M_1\otimes M_2\cdot P$, we have

$$\begin{aligned} f^{(1,1)}(P')&=w_1w_2f^{(1,1)}(P), & f^{(1,2)}(P')&=w_1\lambda_2f^{(1,2)}(P),\\ f^{(2,1)}(P')&=\lambda_1w_2f^{(2,1)}(P), & f^{(2,2)}(P')&=\lambda_1\lambda_2f^{(2,2)}(P). \end{aligned}$$

Let us evaluate these quadratic Markov invariants on a two-leaf phylogenetic tree. In the general case, parameterise the root distribution as $(p_{\mathsf{A}},p_{\mathsf{C}},p_{\mathsf{G}},p_{\mathsf{T}}):=(p+r,q+s,q-s,p-r)$ with $p+q=1$. Immediately after speciation into two taxa we obtain the initial tensor $\tilde{P}$, which, in the standard basis has components:

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

for each $i,j=\mathsf{A},\mathsf{C},\mathsf{G},\mathsf{T}$, and hence in the split basis

$$\begin{pmatrix} \tilde{P}_{00} & \tilde{P}_{01} & \tilde{P}_{00'} & \tilde{P}_{01'}\\ \tilde{P}_{10} & \tilde{P}_{11} & \tilde{P}_{10'} & \tilde{P}_{11'}\\ \tilde{P}_{0'0} & \tilde{P}_{0'1} & \tilde{P}_{0'0'} & \tilde{P}_{0'1'}\\ \tilde{P}_{1'0} & \tilde{P}_{1'1} & \tilde{P}_{1'0'} & \tilde{P}_{1'1'} \end{pmatrix}=\frac{1}{2}\begin{pmatrix} p & 0 & r & 0\\ 0 & q & 0 & s\\ r & 0 & p & 0\\ 0 & s & 0 & q \end{pmatrix}.$$

Extending to two-taxa tree with evolution of taxa 1 and described by strand symmetric transition matrices M_1 and M_2, we obtain the phylogenetic tensor $P=M_1\otimes M_2\cdot\tilde{P}$. However, by a standard argument (the so-called 'pulley-principle' [11]), it is enough to evaluate the special case where after speciation one taxon remains fixed whilst the DNA of the other undergoes random substitutions. Mathematically this allows us to set $M_1\equiv M$ and $M_2=I$. Explicit inspection then shows we have the values $f^{(1,1)}(P)=\frac{1}{4}wpq$, $f^{(1,2)}(P)=\frac{1}{4}wrs$, $f^{(2,1)}(P)=\frac{1}{4}\lambda rs$ and $f^{(2,2)}(P)=\frac{1}{4}\lambda pq$.

Of course 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 2×2 blocks,

$$M = \left(\begin{array}{cc} m & 0 \\ 0 & \overline{m} \end{array} \right).$$

Correspondingly we have the weights $w = \det(m)$ and $\lambda = \det(\overline{m})$. With Jacobi's formula $\det e^{Qt} = e^{tr(Q)t}$ in mind, inspection of the diagonal forms given in Lemma 3 shows that the only generators of the Lie algebra with non-zero trace are $\widehat{R}$ and $\widehat{X}$. Considering the block form and evaluating matrix traces yields

$$\det(m) = e^{-(2\sigma_1+\sigma_2)t}, \qquad \det(\overline{m}) = e^{-\sigma_2 t};$$

where $\sigma_1 := \alpha_3 + \beta_3$ and $\sigma_2 := \alpha_1 + \alpha_2 + \beta_2 + \beta_3$ are the sum of rates *within* and *across* Watson-Crick pairs, respectively. Thus on a two-taxa probability tensor P arising under the strand symmetric model, we have the forms

$$\begin{array}{ll} f^{(1,1)}(P) = \frac{1}{4}e^{-(2\sigma_1-\sigma_2)t}pq, & f^{(1,2)}(P) = \frac{1}{4}e^{-(2\sigma_1+\sigma_2)}rs, \\ f^{(2,1)}(P) = \frac{1}{4}e^{-\sigma_2 t}rs, & f^{(2,2)}(P) = \frac{1}{4}e^{-\sigma_2 t}rs. \end{array}$$

Thus, under the assumption of the strand symmetric model, and depending upon one's willingness to make assumptions about the root distribution parameters p, q, r and s (for example, assume a stationary distribution with $r = s = 0$), it is possible to use the quadratic Markov invariants in a practical setting to obtain independent estimators of the overall within and across rates σ_1 and σ_2.

5. Discussion

In this article we have explored the matrix group properties of the strand symmetric model of DNA evolution from a representation theoretic point of view. We gave a classical decomposition of the Lie algebra associated with the model and further decomposed the representation of this Lie algebra occurring on DNA state space into irreducible modules. We gave a full classification and enumeration of binary and cubic Markov invariants for this model. This work should be seen as distinct from, but complementary to, other results on the strand symmetric model taken from the point of view of "algebraic statistics".

Future work includes the examination of Markov invariants for the strand symmetric model on evolutionary trees with taxa greater than $L = 3$. Of particular interest, are the application of Markov invariants to the quartet case $L = 4$. The case of quartets is of special interest to applied phylogenetics, as this is smallest subset of taxa for which the evolutionary tree history is non-trivial (in the topological sense) relative to Markov models of sequence evolution. Additionally, it is well known that it is enough to recover the evolutionary relations between all quartets of a set of taxa in order to be able to infer the the evolutionary tree of the full set (see [27, Chap. 6] for the relevant discussion). Similarly to the Markov invariants for the general Markov model, as studied in [17], our initial results (unpublished) show that the Markov invariants for the strand symmetric model on quartets of taxa can be used effectively to infer phylogenetic trees.

We defer further speculation on these matters to future work.

References

[1] E. S. Allman and J. A. Rhodes. Phylogenetic ideals and varieties for the general Markov model. *Adv. Appl. Math.*, 40:127–148, 2008.

[2] Elizabeth S. Allman and John A. Rhodes. *Lecture Notes: The Mathematics of Phylogenetics*. IAS/Park City Mathematics Institute, 2005.

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

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

[5] Marta Casanellas and Jesús Fernández-Sánchez. Relevant phylogenetic invariants of evolutionary models. *J. Math. Pures Appl.*, 96:207–229, 2010.

[6] J. A. Cavender and J. Felsenstein. Invariants of phylogenies in a simple case with discrete states. *J. Class.*, 4:57–71, 1987.

[7] J. T. Chang. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Math. Biosci.*, 137(1):51–73, 1996.

[8] Jan Draisma and Jochen Kuttler. On the ideals of equivariant tree models. *Mathematische Annalen*, 344:619–644, 2008.

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

[10] S. N. Evans and T. P. Speed. Invariants of some probability models used in phylogenetic inference. *Ann. Stat.*, 21(1):355–377, 1993.

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

[12] J. Felsenstein. Counting phylogenetic invariants in some simple cases. *J. Theor. Biol.*, 152:357–376, 1991.

[13] J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Sunderland, 2004.

[14] Jesús Fernández-Sánchez, Jeremy G. Sumner, Peter D. Jarvis, and Michael D. Woodhams. Lie Markov models with purine-pyrimidine symmetry. *J. Math. Biol.*, to appear, 2014.

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

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

[17] 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.

[18] 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.

[19] P. D. Jarvis and J.G. Sumner. Adventures in invariant theory. *ANZIAM J.*, 56, in press, 2014.

[20] Steffen Klaere and Volkmar Liebscher. An algebraic analysis of the two state Markov model on tripod trees. *Math. Biosci.*, 237:38–48, 2012.

[21] J. A. Lake. A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony. *Mol. Biol. Evol.*, 4:167–191, 1987.

[22] J. A. Lake. Reconstructing evolutionary trees from DNA and protein sequences: Paralinear distances. *Proceedings of the National Academy of Sciences*, 91:1455–1459, 1994.

[23] D. E. Littlewood. *The Theory of Group Characters*. Clarendon Press, Oxford, 1940.

[24] 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.

[25] I. G. MacDonald. *Symmetric Functions and Hall Polynomials*. Clarendon Press, Oxford, 1979.

[26] Lior Pachter and Bernd Sturmfels, editors. *Algebraic Statistics for Computational Biology*. Cambridge University Press, New York, 2005.

[27] C. Semple and M. Steel. *Phylogenetics*. Oxford University Press, 2003.

[28] B. Sturmfels and S. Sullivant. Toric ideals of phylogenetic invariants. *J. Comput. Biol.*, 12:204–228, 2005.

[29] 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.

[30] J. G. Sumner, J. Fernández-Sánchez, and P. D. Jarvis. Lie Markov models. *J. Theor. Biol.*, 298:16–31, 2012.

[31] 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.

[32] J. G. Sumner and P. D. Jarvis. Markov invariants and the isotropy subgroup of a quartet tree. *J. Theor. Biol.*, 258:302–310, 2009.

[33] 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.

[34] L. A. Székely, M. A. Steel, and P. L. Erdős. Fourier calculus on evolutionary trees. *Adv. Appl. Math.*, 14:200–216, 1993.

[35] S. Tavaré. Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. *Lectures on Mathematics in the Life Sciences (American Society)*, 17:57–86, 1986.

[36] H. Weyl. *The Theory of Groups and Quantum Mechanics*. Dover Publications, 1950.

[37] Wolfram Research, Inc. *Mathematica 8*. Wolfram Research, Inc., Champaign, Illinois, 2010.

[38] Von Bing Yap and Lior Pachter. Identification of evolutionary hotspots in the rodent genomes. *Genome research*, 14(4):574–579, 2004.

APPENDIX A. ENUMERATION OF MARKOV INVARIANTS FOR THE STRAND SYMMETRIC MODEL

The following discussion adopts the notation and adapts the results of [29, 32, 18], and especially [19]. The required background on symmetric function manipulations can be found in the classic text [25].

In the language of representation theory, polynomials in L-way tensors $\psi_{i_1 i_2 \ldots i_L}$ are technically polynomial representations of the underlying matrix groups. For general matrix groups, our 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 $\lambda \vdash m$ with $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_r)$, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r \geq 0$ and $\sum_i \lambda_i = m$. When a partition λ labels directly (and equivalently) a particular gl_n module, irreducible representation, or character, we adopt Littlewood's [23] notation for Schur functions, where the partition is enclosed by curly brackets: $\{\lambda\}$.

For a module $\mathcal{S}$ of a matrix group $G \leq GL(n)$, polynomials of degree D in the components of $\mathcal{S}$ belong to the (in general reducible) module $\mathcal{S} \underline{\otimes} \{D\}$: the plethysm of $\mathcal{S}$ 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 direct product group $G \times G \times \cdots \times G$ (one copy for each leaf on the phylogenetic tree). In this case the resolution of $\big(\otimes^L \mathcal{S} \big) \underline{\otimes} \{D\}$ requires calculation of generic plethysms $\mathcal{S} \underline{\otimes} \sigma$, 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 is taken from [19]:

Lemma 4: General Enumeration of 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_σ occurring in the decomposition of $\mathcal{S} \underline{\otimes} \{\sigma\}$.
2. The number of Markov invariants at degree D is then

$$n_D = \sum_{\sigma_1, \sigma_2, \cdots, \sigma_L \vdash D} g^{(D)}_{\sigma_1 \sigma_2 \cdots \sigma_L} f_{\sigma_1} f_{\sigma_2} \cdots f_{\sigma_L}$$

where $g^{(D)}_{\sigma_1 \sigma_2 \cdots \sigma_L}$ 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, we consider the enumeration of the quadratic, $D = 2$, Markov invariants for the strand symmetric model.

Lemma 5: **Calculation of f_σ for $\mathcal{S}$ at degree $D = 2$.**
At degree $D = 2$ there are two partitions $\sigma \vdash D$ given by (2) and (1^2). Symmetric function manipulations establish that $f_{(1^2)} = 2$ and $f_{(2)} = 1$.

Proof: We appeal to the left-distributive law for plethysms [23]:

$$(A+B)\underline{\otimes}C = \sum_{\mu\subset C} (A\underline{\otimes}\, C/\mu) \otimes (B\underline{\otimes}\mu)\,,$$

where A, B, C are gl_n characters and the summation is over all μ where the skew character C/μ 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 $sl_2 < gl_2$ module and $B \equiv \mathcal{V}$ as a $l_2 < gl_2$ module. We then compute

$$\begin{aligned}(A+B)\,\underline{\otimes}\,\{2\} &= \sum_{\mu=\{0\},\{1\},\{2\}} (A\underline{\otimes}\,\{2\}\,/\mu) \otimes (B\underline{\otimes}\mu)\\ &= (A\underline{\otimes}\,\{2\}) \otimes (B\underline{\otimes}\,\{0\}) + (A\underline{\otimes}\,\{1\}) \otimes (B\underline{\otimes}\,\{1\}) + (A\underline{\otimes}\,\{0\}) \otimes (B\underline{\otimes}\,\{2\})\\ &= A\underline{\otimes}\,\{2\} + A \otimes B + B\underline{\otimes}\,\{2\}\,,\end{aligned}$$

where, in the final line, we have implemented the plethysms $A\underline{\otimes}\,\{1\} = A$ and $B\underline{\otimes}\,\{1\} = B$, and removed the trivial plethysms $A\underline{\otimes}\,\{0\}$ and $B\underline{\otimes}\,\{0\}$ (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 sl_2 module, $A\underline{\otimes}\{2\}$ is irreducible with dimension 3, and similarly considered as a $sl_2 \oplus l_2$ module $A \otimes B$ has dimension $2 \times 2 = 4$ and is irreducible because A is irreducible. However, the general theory in [29] establishes that $B \otimes \{2\}$ contains a one-dimensional submodule of l_2; hence we conclude $f_{(2)} = 1$. A similar calculation establishes

$$(A+B)\,\underline{\otimes}\{1^2\} = A\underline{\otimes}\{1^2\} + A \otimes B + B\underline{\otimes}\{1^2\},$$

and, since both $A\underline{\otimes}\{1^2\}$ and $B\underline{\otimes}\{1^2\}$ 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 $f_{(1^2)} = 2$.

□

In $\mathfrak{S}_2$, (2) is the trivial character and (1^2) is the sgn character. Hence, the "inner" products $(2) * (2) = (2)$, $(2) * (1^2) = (1^2)$, $(1^2) * (1^2) = (2)$ are completely straightforward, and, appealing to associativity, we have

$$g^{(2)}_{\sigma_1\sigma_2\ldots\sigma_L} := \{\text{multiplicity of } (2) \text{ in } \sigma_1 * \sigma_2 * \ldots * \sigma_L\} = \begin{cases} 1, & \text{if } \#\sigma_i = (1^2) \text{ is even,}\\ 0, & \text{otherwise.}\end{cases}$$

Hence applying Lemma 4, we have $n_2 = \sum_{\ell=0}^{\lfloor L/2\rfloor} \binom{L}{2\ell} 2^{2\ell}$ which yields the formula given in Theorem 1 above.

Lemma 6: **Calculation of f_σ for $\mathcal{S}$ at degree $D = 3$.**

At degree $D = 3$ there are two partitions $\sigma \vdash D$ given by (3), $(2,1)$ and (1^3). Symmetric function manipulations establish that $f_{(3)} = f_{(1^3)} = 1$ and $f_{(21)} = 2$.

Proof: With the notation from the previous Lemma, we compute:

$$\begin{aligned}(A+B)\underline{\otimes}\,\{3\} &= \sum_{\mu=\{0\},\{1\},\{2\},\{3\}} (A\underline{\otimes}\,\{3\}\,/\,\{\mu\}) \otimes (B\underline{\otimes}\,\{\mu\})\\ &= A\underline{\otimes}\,\{3\} + (A\underline{\otimes}\,\{2\}) \otimes B + A \otimes (B\underline{\otimes}\,\{2\}) + B\underline{\otimes}\,\{3\}\,.\end{aligned}$$

Similarly, we find:

$$\begin{aligned}(A+B)\,\underline{\otimes}\,\{21\} &= \sum_{\mu=\{0\},\{1\},\{1^2\},\{2\},\{21\}} (A\underline{\otimes}\,\{21\}\,/\,\{\mu\})\otimes(B\underline{\otimes}\,\{\mu\}) \\ &= A\underline{\otimes}\,\{21\} + (A\underline{\otimes}\{2\})\otimes B + \big(A\underline{\otimes}\{1^2\}\big)\otimes B + A\otimes\big(B\underline{\otimes}\{1^2\}\big) \\ &\qquad + A\otimes(B\underline{\otimes}\{2\}) + B\underline{\otimes}\,\{21\}\,,\end{aligned}$$

and

$$\begin{aligned}(A+B)\,\underline{\otimes}\,\{1^3\} &= \sum_{\mu=\{0\},\{1\},\{1^2\},\{1^3\}} \big(A\underline{\otimes}\,\{1^3\}\,/\,\{\mu\}\big)\otimes(B\underline{\otimes}\,\{\mu\}) \\ &= A\underline{\otimes}\,\{1^3\} + \big(A\underline{\otimes}\{1^2\}\big)\otimes B + A\otimes\big(B\underline{\otimes}\{1^2\}\big) + B\underline{\otimes}\{1^3\}.\end{aligned}$$

We identify a single one-dimensional $sl_2 \oplus l_2$ module inside each of $B\underline{\otimes}\,\{3\}$, $\big(A\underline{\otimes}\{1^2\}\big)\otimes B$, $B\underline{\otimes}\,\{21\}$, and $\big(A\underline{\otimes}\{1^2\}\big)\otimes B$. From this we conclude that $f_{(3)} = f_{(1^3)} = 1$ and $f_{(21)} = 2$, as required. □

As characters of $\mathfrak{S}_3$, we have $\sigma = (3), (21)$ or (1^3) and the inner products $(3) * \sigma = \sigma$, $(21) * (21) = (3) + (21) + (1^3)$, $(21) * (1^3) = (21)$ and $(1^3) * (1^3) = (3)$. From these products it is straightforward to establish a recurrence relation for the expansion of $\sigma_1 * \sigma_2 * \ldots * \sigma_L \equiv (3)^i * (21)^j * (1^3)^k = (21)^j * (1^3)^k$ which shows

$$g^{(3)}_{\sigma_1\sigma_2\ldots\sigma_L} = \begin{cases} \frac{1}{3}(2^{k-1} - (-1)^{k-1}), \text{ if } k \neq 0, \\ \frac{1}{2}(1 + (-1)^j), \text{ otherwise.}\end{cases}$$

From this we find that the number of cubic Markov invariants n_3 for the strand symmetric model is given by

$$\begin{aligned} n_3 &= \sum_{\sigma_1,\sigma_2,\ldots,\sigma_L \vdash 3} g^{(3)}_{\sigma_1\sigma_2\ldots\sigma_L} f_{\sigma_1} f_{\sigma_2} \cdots f_{\sigma_L} \\ &= \left(\sum_{k=1}^{L} \binom{L}{k} \frac{1}{3}\left(2^{k-1} - (-1)^{k-1}\right) 2^k\right) + \left(\sum_{j=0}^{L} \binom{L}{j} \frac{1}{2}\left(1 + (-1)^j\right)\right) \\ &= 6^{L-1},\end{aligned}$$

as claimed in Theorem 2.

P D JARVIS, SCHOOL OF MATHEMATICS AND PHYSICS, UNIVERSITY OF TASMANIA, PRIVATE BAG 37, GPO, HOBART TAS 7001, AUSTRALIA

E-mail address: Peter.Jarvis@utas.edu.au

J G SUMNER, SCHOOL OF MATHEMATICS AND PHYSICS, UNIVERSITY OF TASMANIA, PRIVATE BAG 37, GPO, HOBART TAS 7001, AUSTRALIA

E-mail address: Jeremy.Sumner@utas.edu.au