

The Strand Symmetric Model

Marta Casanellas Seth Sullivant

1 Introduction

This chapter is devoted to the study of strand symmetric Markov models on trees from the standpoint of algebraic statistics. By a strand symmetric Markov model, we mean one whose mutation structure reflects the symmetry induced by the double-stranded structure of DNA. In particular, a strand symmetric model for DNA must have the following equalities of probabilities in the root distribution:

$$\pi_A = \pi_T \text{ and } \pi_C = \pi_G$$

and the following equalities of probabilities in the transition matrices (θ_{ij})

$$\theta_{AA} = \theta_{TT}, \theta_{AC} = \theta_{TG}, \theta_{AG} = \theta_{TC}, \theta_{AT} = \theta_{TA},$$

$$\theta_{CA} = \theta_{GT}, \theta_{CC} = \theta_{GG}, \theta_{CG} = \theta_{GC}, \theta_{CT} = \theta_{GA}.$$

Important special cases of strand symmetric Markov models are the group-based phylogenetic models including the Jukes-Cantor model and the Kimura 2 and 3 parameter models. The *general strand symmetric model* or in this chapter just the *strand symmetric model* (SSM) has only these eight equalities of probabilities in the transition matrices and no further restriction on the transition probabilities. Thus, for each edge in the corresponding phylogenetic model, there are 6 free parameters.

For the standard group-based models (i.e. Jukes-Cantor and Kimura), the transition matrices and the entire parametrization can be simultaneously diagonalized by means of the Fourier transform of the group $\mathbb{Z}_2 \times \mathbb{Z}_2$ [2, 6]. Besides the practical uses of the Fourier transform for group based models (see for example [4]), this diagonalization of the group-based models makes it possible to compute phylogenetic invariants for these models, by reducing

the problem to the claw tree $K_{1,3}$ [5]. Our goal in this chapter is to extend the Fourier transform from group-based models to the strand symmetric model. This is carried out in Section 2.

In Section 3 we focus in on the case of the three taxa tree. The computation of phylogenetic invariants for the SSM in the Fourier coordinates is still not complete, though we report on what is known about these invariants. In particular, we describe all invariants of degree three and four. Section 5 is concerned with extending known invariants from the three taxa tree to an arbitrary tree. In particular, we describe how to extend the given degree three and four invariants from Section 3 to an arbitrary binary tree. To do this, we introduce *G-tensors* and explore their properties in Section 4.

In Section 6, we take up the task of extending the “gluing” results for phylogenetic invariants which appear both in the work of Allman and Rhodes [1] and Sturmfels and Sullivant [5]. Our exposition and inspiration mainly comes from the work of Allman and Rhodes and we deduce that the problem of determining *defining* phylogenetic invariants for the strand symmetric model reduces to finding phylogenetic invariants for the claw tree $K_{1,3}$. Here *defining* means a set of polynomials which generate the ideal of invariants up to radical; that is, defining invariants have the same zero set as the whole ideal of invariants. This result is achieved by proving some “block diagonal” versions of results which appear in the Allman and Rhodes paper. This line of attack is the heart of Sections 4 and 6.

2 Matrix-Valued Fourier Transform

In this section we introduce the matrix-valued group-based models and show that the strand symmetric model is a matrix-valued group-based model. Then we describe the matrix-valued Fourier transform and the resulting simplification in the parametrization of these models. We make special emphasis on the strand symmetric model.

Let T be a rooted tree with n taxa. First, we wish to describe the random variables associated to each vertex in the tree in the matrix-valued group-based models. Each random variable X_v takes on kl states where k is the cardinality of a finite abelian group G . The states of the random variable are 2-ples $\binom{j}{i}$ where $j \in G$ and $i \in \{0, 1, \dots, l-1\}$.

Associated to the root node R in the tree is the root distribution $R_{i_1}^{j_1}$. For each edge E of T the double indexed set of parameters $E_{i_1 i_2}^{j_1 j_2}$ is the transition

matrix associated to this edge. We use the convention that E is both the edge and the transition matrix associated to that edge, to avoid the need for introducing a third index on the matrices. Thus $E_{i_1 i_2}^{j_1 j_2}$ is the conditional probability of making a transition from state $\binom{j_1}{i_1}$ to state $\binom{j_2}{i_2}$ along the edge E .

Definition 1. A phylogenetic model is a *matrix-valued group-based model* if for each edge, the matrix transition probabilities satisfy

$$E_{i_1 i_2}^{j_1 j_2} = E_{i_1 i_2}^{j_3 j_4}$$

when $j_1 - j_2 = j_3 - j_4$ (where the difference is taken in G) and the root distribution probabilities satisfy $R_i^{j_1} = R_i^{j_2}$.

Example 2. Consider the strand symmetric model and make the identification of the states $A = \binom{0}{0}$, $G = \binom{0}{1}$, $T = \binom{1}{0}$, and $C = \binom{1}{1}$. One can check directly from the definitions that the strand symmetric model is a matrix-valued group-based model with $l = 2$ and $G = \mathbb{Z}_2$.

To avoid some even more cumbersome notation, we will restrict attention to binary trees T and to the strand symmetric model for DNA. While the results of Section 3 and 5 are exclusive to the case of the SSM, all our other results can be easily extended to arbitrary matrix-valued group-based models with the introduction of the more general Fourier transform, though we will not explain these generalizations here.

We assume all edges of T are directed away from the root R . Given an edge E of T let $s(E)$ denote the initial vertex of E and $t(E)$ the trailing vertex. Then the parametrization of the phylogenetic model is given as follows. The probability of observing states $\binom{j_1 j_2 \dots j_n}{i_1 i_2 \dots i_n}$ at the leaves is

$$P_{i_1 i_2 \dots i_n}^{j_1 j_2 \dots j_n} = \sum_{\left(\binom{j_v}{i_v}\right) \in H} R_{i_R}^{j_R} \prod_E E_{i_{s(E)} i_{t(E)}}^{j_{s(E)} j_{t(E)}}$$

where the product is taken over all edges E of T and the sum is taken over the set

$$H = \left\{ \left(\binom{j_v}{i_v} \right)_{v \in \text{Int}V(T)} \mid j_v, i_v \in \{0, 1\} \right\}.$$

Here $\text{Int}V(T)$ denotes the interior or nonleaf vertices of T .

Example 3. For the three leaf claw tree, the parametrization is given by the expression:

$$P_{ijk}^{lmn} = R_0^0 A_{0i}^{0l} B_{0j}^{0m} C_{0k}^{0n} + R_0^1 A_{0i}^{1l} B_{0j}^{1m} C_{0k}^{1n} + R_1^0 A_{1i}^{0l} B_{1j}^{0m} C_{1k}^{0n} + R_1^1 A_{1i}^{1l} B_{1j}^{1m} C_{1k}^{1n}.$$

The study of this particular tree will occupy a large part of the paper.

Because of the role of the group in determining the symmetry in the parametrization, the Fourier transform can be applied to make the parametrization simpler. We will not define the Fourier transform in general, only in the specific case of the group \mathbb{Z}_2 . The Fourier transform applies to all of the probability coordinates, the transition matrices and the root distribution.

Definition 4. The *Fourier transform of the probability coordinates* is

$$q_{i_1 i_2 \dots i_n}^{j_1 j_2 \dots j_n} = \sum_{k_1, k_2, \dots, k_n \in \{0,1\}} (-1)^{k_1 j_1 + k_2 j_2 + \dots + k_n j_n} p_{i_1 i_2 \dots i_n}^{k_1 k_2 \dots k_n}.$$

The *Fourier transform of the transition matrix* E is

$$e_{i_1 i_2}^{j_1 j_2} = \frac{1}{2} \sum_{k_1, k_2 \in \{0,1\}} (-1)^{k_1 j_1 + k_2 j_2} E_{i_1 i_2}^{k_1 k_2}.$$

The *Fourier transform of the root distribution* is

$$r_i^j = \sum_{k \in \{0,1\}} (-1)^{kj} R_i^k.$$

It is easy to check that $e_{i_1 i_2}^{j_1 j_2} = 0$ if $j_1 + j_2 = 1 \in \mathbb{Z}_2$ and similarly that $r_i^j = 0$ if $j = 1$. In particular, writing e as a matrix, we see that the Fourier transform replaces the matrix E with a matrix e that is block diagonal. Generally, when working with our “hands on” the parameters (in particular in Section 3) we will write the transition matrices with only one superscript: $e_{i_1 i_2}^j$ and the transformed root distribution r_i with no superscript at all, though at other times it will be more convenient to have the extra superscript around, in spite of their redundancy.

Lemma 5. In the *Fourier coordinates* the parametrization is given by the rule

$$q_{i_1 i_2 \dots i_n}^{j_1 j_2 \dots j_n} = \sum_{(i_v) \in H} r_{i_r}^{j_r} \prod_e e_{i_{s(e)} i_{t(e)}}^{j_{s(e)}}$$

where $j_{s(e)}$ is the sum of j_l such that l is a leaf below $s(e)$ in the tree, $j_r = j_1 + \dots + j_n$ and H denotes the set

$$H = \{(i_v)_{v \in \text{Int}V(T)} \text{ and } i_v \in \{0, 1\}\}.$$

Proof. We can rewrite the parametrization in the probability coordinates as

$$p_{i_1 i_2 \dots i_n}^{k_1 k_2 \dots k_n} = \sum_{(i_v) \in H} \left(\sum_{(k_v) \in H'} R_{i_R}^{k_R} \prod_E E_{i_{s(E)} i_{t(E)}}^{k_{s(E)} k_{t(E)}} \right)$$

where H is the set defined in the lemma and

$$H' = \{(k_v)_{v \in \text{Int}V(T)} \text{ and } k_v \in \mathbb{Z}_2\}.$$

The crucial observation is that for any fixed values of i_1, \dots, i_n and $(i_v) \in H$, the expression inside the parentheses is a standard group-based model for \mathbb{Z}_2 . Applying the Fourier transform we have

$$q_{i_1 i_2 \dots i_n}^{j_1 j_2 \dots j_n} = \sum_{k_1, k_2, \dots, k_n \in \{0, 1\}} (-1)^{k_1 j_1 + k_2 j_2 + \dots + k_n j_n} \sum_{(i_v) \in H} \left(\sum_{(k_v) \in H'} R_{i_R}^{k_R} \prod_E E_{i_{s(E)} i_{t(E)}}^{k_{s(E)} k_{t(E)}} \right)$$

and interchanging summations

$$= \sum_{(i_v) \in H} \left(\sum_{k_1, k_2, \dots, k_n \in \{0, 1\}} (-1)^{k_1 j_1 + k_2 j_2 + \dots + k_n j_n} \sum_{(k_v) \in H'} R_{i_R}^{k_R} \prod_E E_{i_{s(E)} i_{t(E)}}^{k_{s(E)} k_{t(E)}} \right).$$

By our crucial observation above, the expression inside the large parentheses is the Fourier transform of a group-based model and hence by results in [2] and [6] the expression inside the parentheses factors in terms of the Fourier transforms of the transition matrices and root distribution in precisely the way illustrated in the statement of the lemma. \square

Definition 6. Given a tree T , the projective variety of the SSM given by the tree T is denoted by $V(T)$. The notation $CV(T)$ denotes the affine cone over $V(T)$.

Proposition 7 (Linear Invariants).

$$q_{i_1 i_2 \dots i_n}^{j_1 j_2 \dots j_n} = 0$$

if $j_1 + j_2 + \dots + j_n = 1 \in \mathbb{Z}_2$.

Proof. The equation $j_1 + j_2 + \dots + j_n = 1 \in \mathbb{Z}_2$ implies that in the parametrization every summand involves $r_{i_r}^1$ for some i_r . However, all of these parameters are zero. \square

The linear invariants in the previous lemma are equivalent to the fact that

$$p_{i_1 i_2 \dots i_n}^{j_1 j_2 \dots j_n} = p_{i_1 i_2 \dots i_n}^{\bar{j}_1 \bar{j}_2 \dots \bar{j}_n}$$

where $\bar{j} = 1 - j \in \mathbb{Z}_2$.

Up until now, we have implicitly assumed that all the matrices E involved were actually matrices of transition probabilities and that the root distribution R was an honest probability distribution. If we drop these conditions and look at the parametrization in the Fourier coordinates, we can, in fact, drop r from this representation altogether. That is, the variety parametrized by dropping the transformed root distribution r is the cone over the Zariski closure of the probabilistic parametrization.

Lemma 8. *In the Fourier coordinates there is an open subset of $CV(T)$, the cone over the strand symmetric model, that can be parametrized as*

$$q_{i_1 i_2 \dots i_n}^{j_1 j_2 \dots j_n} = \sum_{(i_v) \in H} \prod_e e_{i_{s(e)} i_{t(e)}}^{j_{s(e)}}$$

when $j_1 + \dots + j_n = 0 \in \mathbb{Z}_2$, where $j_{s(e)}$ is the sum of j_l such that l is a leaf below $s(e)$ in the tree and H denotes the set

$$H = \{(i_v)_{v \in \text{Int}V(T)} \text{ and } i_v \in \{0, 1\}\}.$$

Proof. Due to the structure of the reparametrization of the SSM which we will prove in Section 4, it suffices to prove the lemma when T is the 3-leaf claw tree $K_{1,3}$. In this case, we are comparing the parametrizations

$$\phi : q_{ijk}^{mno} = r_0 a_{0i}^m b_{0j}^n c_{0k}^o + r_1 a_{1i}^m b_{1j}^n c_{1k}^o$$

and

$$\psi : q_{ijk}^{mno} = a_{0i}^m e_{0j}^n f_{0k}^o + a_{1i}^m e_{1j}^n f_{1k}^o.$$

In the second case, there are no conditions on the parameters. In the first parametrization, the stochastic assumption on the root distribution and transition matrices translates into the following restrictions on the Fourier parameters

$$r_0 = 1, a_{i0}^0 + a_{i1}^0 = 1, b_{i0}^0 + b_{i1}^0 = 1, c_{i0}^0 + c_{i1}^0 = 1$$

for $l = 0, 1$. We must show that for d, e, f belonging to some open subset U we can choose r, a, b, c with the prescribed restrictions which realize the same Q tensor up to scaling. To do this, define

$$\delta_l = d_{l0}^0 + d_{l1}^0, \gamma_l = e_{l0}^0 + e_{l1}^0, \lambda_l = f_{l0}^0 + f_{l1}^0$$

for $l = 0, 1$ and take U the subset where these numbers are all non-zero. Set

$$a_{li}^m = \delta_l^{-1} d_{li}^m, b_{lj}^n = \gamma_l^{-1} e_{lj}^n, c_{lk}^o = \lambda_l^{-1} f_{lk}^o, r_0 = 1, \text{ and } r_1 = \frac{\delta_1 \gamma_1 \lambda_1}{\delta_0 \gamma_0 \lambda_0}.$$

Clearly, all the parameters r, a, b, c satisfy the desired prescription. Furthermore, the parameterization with this choice of r, a, b, c differs from the original parametrization by a factor of $(\delta_0 \gamma_0 \lambda_0)^{-1}$. This proves that $\psi(U) \subset \text{Im}(\psi) \subset CV(T)$. On the other hand we have that $V(T) \subset \text{Im}(\psi)$ because we can always take $d_{li}^m = r_l a_{li}^m, e_{lj}^n = b_{lj}^n, f_{lk}^o = c_{lk}^o$. Moreover it is clear that $\text{Im}(\psi)$ is a cone and hence $CV(T) \subset \text{Im}(\psi)$. The proof of the lemma is completed by taking the Zariski closure. \square

Example 9. In the particular instance of the three leaf claw tree the Fourier parametrization of the model is given by the formula

$$q_{ijk}^{mno} = a_{0i}^m b_{0j}^n c_{0k}^o + a_{1i}^m b_{1j}^n c_{1k}^o.$$

3 Invariants for the 3 taxa tree

In this section, we will describe the degree 3 and degree 4 phylogenetic invariants for the claw tree $K_{1,3}$ on the strand symmetric model. We originally found these polynomial invariants using the computational algebra package Macaulay2 [3] though we will give a combinatorial description of these invariants and proofs that they do, in fact, vanish on the strand symmetric model.

It is still an open problem to decide whether or not the 32 cubics and 18 quartics described here generate the ideal of invariants, or even describe the SSM set theoretically. Computationally, we determined that they generate the ideal up to degree 4. Furthermore, one can show that neither the degree 3 nor the degree 4 invariants alone are sufficient to describe the variety set theoretically.

3.1 Degree 3 Invariants

Proposition 10. For each $l = 1, 2, 3$ let $m_l, n_l, o_l, i_l, j_l, k_l$ be indices in $\{0, 1\}$ such that $m_l + n_l + o_l = 0$, $m_1 = m_2$, $m_3 = 1 - m_1$, $n_1 = n_3$, $n_2 = 1 - n_1$, $o_2 = o_3$, and $o_1 = 1 - o_2$ in \mathbb{Z}_2 . Let $f(m_\bullet, n_\bullet, o_\bullet, i_\bullet, j_\bullet, k_\bullet)$ be the polynomial in the Fourier coordinates described as

$$\begin{vmatrix} q_{i_1 j_1 k_1}^{m_1 n_1 o_1} & q_{i_2 j_1 k_1}^{m_2 n_1 o_1} & 0 \\ q_{i_1 j_2 k_2}^{m_1 n_2 o_2} & q_{i_2 j_2 k_2}^{m_2 n_2 o_2} & q_{i_3 j_3 k_2}^{m_3 n_3 o_2} \\ q_{i_1 j_2 k_3}^{m_1 n_2 o_3} & q_{i_2 j_2 k_3}^{m_2 n_2 o_3} & q_{i_3 j_3 k_3}^{m_3 n_3 o_3} \end{vmatrix} - \begin{vmatrix} q_{i_1 j_3 k_1}^{m_1 n_3 o_1} & q_{i_2 j_3 k_1}^{m_2 n_3 o_1} & 0 \\ q_{i_1 j_2 k_2}^{m_1 n_2 o_2} & q_{i_2 j_2 k_2}^{m_2 n_2 o_2} & q_{i_3 j_1 k_2}^{m_3 n_1 o_2} \\ q_{i_1 j_2 k_3}^{m_1 n_2 o_3} & q_{i_2 j_2 k_3}^{m_2 n_2 o_3} & q_{i_3 j_1 k_3}^{m_3 n_1 o_3} \end{vmatrix}.$$

Then $f(m_\bullet, n_\bullet, o_\bullet, i_\bullet, j_\bullet, k_\bullet)$ is a phylogenetic invariant for $K_{1,3}$ on the SSM.

Remark 11. The only nonzero cubics invariants for $K_{1,3}$ arising from Proposition 1 are those satisfying $i_2 = 1 - i_1$, $i_3 = i_2, j_2 = j_1$, $j_3 = 1 - j_1$, $k_2 = 1 - k_1$ and $k_3 = k_1$. We maintain all the indices because they are necessary when we extend invariants to larger trees in section 5. In total, we obtain 32 invariants in this way and we verified in Macaulay2 [3] that these 32 invariants generate the ideal in degree 3.

Proof. In order to prove this result it is very useful to write the parametrization in Fourier coordinates as:

$$q_{ijk}^{mno} = \begin{vmatrix} a_{0i}^m b_{0j}^n & -c_{1k}^o \\ a_{1i}^m b_{1j}^n & c_{0k}^o \end{vmatrix}.$$

In $f(m_\bullet, n_\bullet, o_\bullet, i_\bullet, j_\bullet, k_\bullet)$ we substitute the Fourier coordinates by their parametrization and we call D_1 the first determinant and D_2 the second one so that

$$D_1 = \begin{vmatrix} \begin{vmatrix} a_{0i_1}^{m_1} b_{0j_1}^{n_1} & -c_{1k_1}^{o_1} \\ a_{1i_1}^{m_1} b_{1j_1}^{n_1} & c_{0k_1}^{o_1} \end{vmatrix} & \begin{vmatrix} a_{0i_2}^{m_2} b_{0j_1}^{n_1} & -c_{1k_1}^{o_1} \\ a_{1i_2}^{m_2} b_{1j_1}^{n_1} & c_{0k_1}^{o_1} \end{vmatrix} & 0 \\ \begin{vmatrix} a_{0i_1}^{m_1} b_{0j_2}^{n_2} & -c_{1k_2}^{o_2} \\ a_{1i_1}^{m_1} b_{1j_2}^{n_2} & c_{0k_2}^{o_2} \end{vmatrix} & \begin{vmatrix} a_{0i_2}^{m_2} b_{0j_2}^{n_2} & -c_{1k_2}^{o_2} \\ a_{1i_2}^{m_2} b_{1j_2}^{n_2} & c_{0k_2}^{o_2} \end{vmatrix} & \begin{vmatrix} a_{0i_3}^{m_3} b_{0j_3}^{n_3} & -c_{1k_2}^{o_2} \\ a_{1i_3}^{m_3} b_{1j_3}^{n_3} & c_{0k_2}^{o_2} \end{vmatrix} \\ \begin{vmatrix} a_{0i_1}^{m_1} b_{0j_2}^{n_2} & -c_{1k_3}^{o_3} \\ a_{1i_1}^{m_1} b_{1j_2}^{n_2} & c_{0k_3}^{o_3} \end{vmatrix} & \begin{vmatrix} a_{0i_2}^{m_2} b_{0j_2}^{n_2} & -c_{1k_3}^{o_3} \\ a_{1i_2}^{m_2} b_{1j_2}^{n_2} & c_{0k_3}^{o_3} \end{vmatrix} & \begin{vmatrix} a_{0i_3}^{m_3} b_{0j_3}^{n_3} & -c_{1k_3}^{o_3} \\ a_{1i_3}^{m_3} b_{1j_3}^{n_3} & c_{0k_3}^{o_3} \end{vmatrix} \end{vmatrix}.$$

Now we observe that the indices in the first position are the same for each column in both determinants involved in $f(m_\bullet, n_\bullet, o_\bullet, i_\bullet, j_\bullet, k_\bullet)$. Similarly, the

indices in the third position are the same for each row in both determinants. Using recursively the formula

$$\left| \begin{array}{c|c|c|c|} \left| \begin{array}{cc} x_{1,1} & y \\ x_{2,1} & z \end{array} \right| & \left| \begin{array}{cc} x_{1,2} & y \\ x_{2,2} & x \end{array} \right| & \cdots & \left| \begin{array}{cc} x_{1,n} & y \\ x_{2,n} & z \end{array} \right| \\ \hline x_{3,1} & x_{3,2} & \cdots & x_{3,n} \\ \hline \vdots & \vdots & \vdots & \vdots \\ \hline x_{n+1,1} & x_{n+1,2} & \cdots & x_{n+1,n} \end{array} \right| = \left| \begin{array}{ccccc} x_{1,1} & x_{1,2} & \cdots & x_{1,n} & y \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} & z \\ x_{3,1} & x_{3,2} & \cdots & x_{3,n} & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n+1,1} & x_{n+1,2} & \cdots & x_{n+1,n} & 0 \end{array} \right|$$

it is easy to see that D_1 can be written as the following 6×6 determinant:

$$D_1 = \begin{vmatrix} a_{0i_1}^{m_1} b_{0j_1}^{n_1} & a_{0i_2}^{m_2} b_{0j_1}^{n_1} & 0 & -c_{1k_1}^{o_1} & 0 & 0 \\ a_{1i_1}^{m_1} b_{1j_1}^{n_1} & a_{1i_2}^{m_2} b_{1j_1}^{n_1} & 0 & c_{0k_1}^{o_1} & 0 & 0 \\ a_{0i_1}^{m_1} b_{0j_2}^{n_2} & a_{0i_2}^{m_2} b_{0j_2}^{n_2} & a_{0i_3}^{m_3} b_{0j_3}^{n_3} & 0 & -c_{1k_2}^{o_2} & 0 \\ a_{1i_1}^{m_1} b_{1j_2}^{n_2} & a_{1i_2}^{m_2} b_{1j_2}^{n_2} & a_{1i_3}^{m_3} b_{1j_3}^{n_3} & 0 & c_{0k_2}^{o_2} & 0 \\ a_{0i_1}^{m_1} b_{0j_2}^{n_2} & a_{0i_2}^{m_2} b_{0j_2}^{n_2} & a_{0i_3}^{m_3} b_{0j_3}^{n_3} & 0 & 0 & -c_{1k_3}^{o_3} \\ a_{1i_1}^{m_1} b_{1j_2}^{n_2} & a_{1i_2}^{m_2} b_{1j_2}^{n_2} & a_{1i_3}^{m_3} b_{1j_3}^{n_3} & 0 & 0 & c_{0k_3}^{o_3} \end{vmatrix}.$$

Now using Laplace expansion for the last 3 columns we see that D_1 is equal to

$$\begin{aligned} & -c_{0k_1}^{o_1} c_{1k_2}^{o_2} c_{0k_3}^{o_3} \begin{vmatrix} a_{0i_1}^{m_1} b_{0j_1}^{n_1} & a_{0i_2}^{m_2} b_{0j_1}^{n_1} & 0 \\ a_{1i_1}^{m_1} b_{1j_2}^{n_2} & a_{1i_2}^{m_2} b_{1j_2}^{n_2} & a_{1i_3}^{m_3} b_{1j_3}^{n_3} \\ a_{0i_1}^{m_1} b_{0j_2}^{n_2} & a_{0i_2}^{m_2} b_{0j_2}^{n_2} & a_{0i_3}^{m_3} b_{0j_3}^{n_3} \end{vmatrix} - c_{0k_1}^{o_1} c_{0k_2}^{o_2} c_{1k_3}^{o_3} \begin{vmatrix} a_{0i_1}^{m_1} b_{0j_1}^{n_1} & a_{0i_2}^{m_2} b_{0j_1}^{n_1} & 0 \\ a_{0i_1}^{m_1} b_{0j_2}^{n_2} & a_{0i_2}^{m_2} b_{0j_2}^{n_2} & a_{0i_3}^{m_3} b_{0j_3}^{n_3} \\ a_{1i_1}^{m_1} b_{1j_2}^{n_2} & a_{1i_2}^{m_2} b_{1j_2}^{n_2} & a_{1i_3}^{m_3} b_{1j_3}^{n_3} \end{vmatrix} + \\ & c_{1k_1}^{o_1} c_{1k_2}^{o_2} c_{0k_3}^{o_3} \begin{vmatrix} a_{1i_1}^{m_1} b_{1j_1}^{n_1} & a_{1i_2}^{m_2} b_{1j_1}^{n_1} & 0 \\ a_{1i_1}^{m_1} b_{1j_2}^{n_2} & a_{1i_2}^{m_2} b_{1j_2}^{n_2} & a_{1i_3}^{m_3} b_{1j_3}^{n_3} \\ a_{0i_1}^{m_1} b_{0j_2}^{n_2} & a_{0i_2}^{m_2} b_{0j_2}^{n_2} & a_{0i_3}^{m_3} b_{0j_3}^{n_3} \end{vmatrix} + c_{1k_1}^{o_1} c_{0k_2}^{o_2} c_{1k_3}^{o_3} \begin{vmatrix} a_{1i_1}^{m_1} b_{1j_1}^{n_1} & a_{1i_2}^{m_2} b_{1j_1}^{n_1} & 0 \\ a_{0i_1}^{m_1} b_{0j_2}^{n_2} & a_{0i_2}^{m_2} b_{0j_2}^{n_2} & a_{0i_3}^{m_3} b_{0j_3}^{n_3} \\ a_{1i_1}^{m_1} b_{1j_2}^{n_2} & a_{1i_2}^{m_2} b_{1j_2}^{n_2} & a_{1i_3}^{m_3} b_{1j_3}^{n_3} \end{vmatrix} \end{aligned}$$

Doing the same procedure for D_2 we see that its Laplace expansion has exactly the same 4 nonzero terms. \square

3.2 Degree 4 Invariants

Now we wish to explain the derivation of some nontrivial degree 4 invariants for the SSM on the $K_{1,3}$ tree. Each of the degree 4 invariants involves 16 of the nonzero Fourier coordinates which come from choosing two possible

distinct sets of indices for the group-based indices. Up to symmetry, we may suppose these are from the tensors

$$q^{mn_1o_1} \text{ and } q^{mn_2o_2}.$$

Choose the ten indices $i_1, i_2, j_1, j_2, j_3, j_4, k_1, k_2, k_3, k_4$. Define the four matrices $q_i^{mn_1o_1}$ and $q_i^{mn_2o_2}$, $i \in \{i_1, i_2\}$ by

$$q_i^{mn_1o_1} = \begin{pmatrix} q_{ij_1k_1}^{mn_1o_1} & q_{ij_1k_2}^{mn_1o_1} \\ q_{ij_2k_1}^{mn_1o_1} & q_{ij_2k_2}^{mn_1o_1} \end{pmatrix} \text{ and } q_i^{mn_2o_2} = \begin{pmatrix} q_{ij_3k_3}^{mn_2o_2} & q_{ij_3k_4}^{mn_2o_2} \\ q_{ij_4k_3}^{mn_2o_2} & q_{ij_4k_4}^{mn_2o_2} \end{pmatrix}$$

For any of these matrices, adding an extra subindex j means taking the j -th row of the matrix, e.g. q_{1j}^{011} is the vector $(q_{1j0}^{011} \ q_{1j1}^{011})$.

Theorem 12. *The 2×2 minors of the following 2×3 matrix are all degree 4 invariants of the SSM model on the 3 leaf claw tree:*

$$\begin{pmatrix} |q_{i_1}^{mn_1o_1}| & \begin{vmatrix} q_{i_1j_1}^{mn_1o_1} \\ q_{i_2j_2}^{mn_1o_1} \end{vmatrix} & + & \begin{vmatrix} q_{i_1j_2}^{mn_1o_1} \\ q_{i_2j_1}^{mn_1o_1} \end{vmatrix} & |q_{i_2}^{mn_1o_1}| \\ |q_{i_1}^{mn_2o_2}| & \begin{vmatrix} q_{i_1j_3}^{mn_2o_2} \\ q_{i_2j_4}^{mn_2o_2} \end{vmatrix} & + & \begin{vmatrix} q_{i_1j_4}^{mn_2o_2} \\ q_{i_2j_3}^{mn_2o_2} \end{vmatrix} & |q_{i_2}^{mn_2o_2}| \end{pmatrix}.$$

These degree 4 invariants are not in the radical of the ideal generated by the degree 3 invariants above. Up to symmetry, of the SSM on $K_{1,3}$ the 18 degree 4 invariants which arise this way are the only minimal generators of the ideal of degree 4.

Proof. The third claim was proven computationally using Macaulay 2 [3]. The second claim follows by noting that all of the degree 3 invariants above use 3 different superscripts whereas the degree 4 invariants use only 2 different superscripts and the polynomials are multi-homogeneous in these indices. Hence, for example, an assignment of arbitrary values to the tensors q^{000} and q^{011} and setting $q^{101} = q^{110} = 0$ creates a set of Fourier values which necessarily satisfies all degree 3 invariants but does not satisfy the degree 4 polynomials described in the statement of the theorem.

Now we will prove that these polynomials are, in fact, invariants of the SSM on $K_{1,3}$. The parametrization of $q_i^{mn_1o_1}$ and $q_i^{mn_2o_2}$ can be rewritten as

$$q_i^{mn_1o_1} = a_{0i}^m M_0^0 + a_{1i}^m M_1^0 \text{ and } q_i^{mn_2o_2} = a_{0i}^m M_0^1 + a_{1i}^m M_1^1$$

where each of the four matrices M_0^0 , M_1^0 , M_0^1 , and M_1^1 are arbitrary 2×2 matrices of rank 1. This follows by noting that the $q^{mn_1o_1}$ uses b^{n_1} and c^{o_1} in its description, $q^{mn_2o_2}$ uses b^{n_2} and c^{o_2} in its description, and simply reforming these descriptions into matrix notation.

In particular, $q_{i_1}^{mn_1o_1}$ and $q_{i_2}^{mn_1o_1}$ lie in the plane spanned by the rank 1 matrices M_0^0 and M_1^0 and $q_{i_1}^{mn_2o_2}$ and $q_{i_2}^{mn_2o_2}$ lie in the plane spanned by the rank 1 matrices M_0^1 and M_1^1 . Furthermore, the coefficients used to write these linear combinations are the same for pair $q_{i_1}^{mn_1o_1}$ and $q_{i_1}^{mn_2o_2}$ and for the pair $q_{i_2}^{mn_1o_1}$ and $q_{i_2}^{mn_2o_2}$.

If we are given a general point on the variety of the SSM, none of the matrices $q_{i_1}^{mn_1o_1}$, $q_{i_2}^{mn_1o_1}$, $q_{i_1}^{mn_2o_2}$ or $q_{i_2}^{mn_2o_2}$ will have rank 1. This implies that, generically, the set of matrices

$$\mathcal{M}_1 = \{\lambda q_{i_1}^{mn_1o_1} + \gamma q_{i_2}^{mn_1o_1} | \lambda, \gamma \in \mathbb{C}\}$$

$$\mathcal{M}_2 = \{\lambda q_{i_1}^{mn_2o_2} + \gamma q_{i_2}^{mn_2o_2} | \lambda, \gamma \in \mathbb{C}\}$$

each contain precisely 2 lines of rank 1 matrices. This is because the variety of 2×2 rank 1 matrices has degree 2. The set of values of λ and γ which produce these lines of rank 1 matrices are the same because of the way that $q_{i_1}^{mn_1o_1}$, $q_{i_2}^{mn_1o_1}$, $q_{i_1}^{mn_2o_2}$ or $q_{i_2}^{mn_2o_2}$ were written in terms of M_0^0 , M_1^0 , M_0^1 and M_1^1 . In the first case, this set of λ and γ is the solution set of the quadratic equation

$$|\lambda q_{i_1}^{mn_1o_1} + \gamma q_{i_2}^{mn_1o_1}| = 0$$

and in the second case this set is the solution to the quadratic equation

$$|\lambda q_{i_1}^{mn_2o_2} + \gamma q_{i_2}^{mn_2o_2}| = 0.$$

To say that these two quadrics have the same zero set is equivalent to the vanishing of the three 2×2 minors in the statement of the theorem. Since the minors in the statement of the theorem vanish for a general point on the parametrization they must vanish on the entire variety and hence are invariants of the SSM. \square

4 G -tensors

In this section we introduce the notion of a G -tensor which should be regarded as a multidimensional analog of a block diagonal matrix. We describe G -tensor multiplication and a certain variety defined for G -tensors which will

be useful for extending invariants from the 3-leaf claw tree to arbitrary trees. This variety ${}^G V(4^{r_1}, 4^{r_2}, 4^{r_3})$ generalizes in a natural way the SSM on the claw tree $K_{1,3}$.

Notation. Let G be a group. For an n -tuple $\mathbf{j} = (j_1, \dots, j_n) \in G^n$ we denote by $\sigma(\mathbf{j})$ the sum $j_1 + \dots + j_n \in G$.

Definition 13. Let $q_{i_1 \dots i_n}^{\mathbf{j}_1 \dots \mathbf{j}_n}$ define a $4^{r_1} \times \dots \times 4^{r_n}$ tensor Q where the upper indices are r_i -tuples in the group $G = \mathbb{Z}_2$. We say that Q is G -tensor if whenever $\sigma(\mathbf{j}_1) + \dots + \sigma(\mathbf{j}_n) \neq 0$ in G , $q_{i_1 \dots i_n}^{\mathbf{j}_1 \dots \mathbf{j}_n} = 0$. If $n = 2$ then Q is called a G -matrix.

Lemma 14. *If Q is a $4 \times \dots \times 4$ tensor arising from the SSM in the Fourier coordinates, then Q is a G -tensor. All the Fourier parameter matrices $e_{i_1 i_2}^{j_1 j_2}$ are G -matrices.*

Proof. This is immediate from Proposition 7 and the comments following Definition 4. \square

Convention. Henceforth, we order any set of indices $\left\{ \binom{j_1 \dots j_t}{i_1 \dots i_t} \right\}_{j_1, \dots, j_t, i_1, \dots, i_t}$ so that we put first those indices whose upper sum $\sigma(\mathbf{j}) = j_1 + \dots + j_t$ is equal to zero.

From now on we are going to use only Fourier coordinates and we will refer to the corresponding tensor as Q .

An operation on tensors that we will use frequently is the tensor multiplication $*$ which is defined as follows. If R and Q are n -dimensional and m -dimensional tensors so that R (resp. Q) has κ states at the last index (resp. first index), the $(m + n - 2)$ -dimensional tensor $R * Q$ is defined as

$$(R * Q)_{i_1, \dots, i_{n+m-2}} = \sum_{j=1}^{\kappa} R_{i_1, \dots, i_{n-1}, j} \cdot Q_{j, i_n, \dots, i_{n+m-2}}.$$

If R and Q are matrices this is the usual matrix multiplication. Note that if R and Q are G -tensors then $R * Q$ is also a G -tensor. We can also perform the $*$ operation on two varieties: if V and W are varieties of tensors then $V * W = \overline{\{R * Q \mid R \in V, Q \in W\}}$. If T' is a tree with taxa v_1, \dots, v_n and T'' is a tree with taxa w_1, \dots, w_m we call $T' * T''$ the tree obtained by identifying the vertices v_n and w_1 , deleting this new vertex, and replacing the two corresponding edges by a single edge. This construction is a useful tool for constructing a reparametrization of the variety associated to an n -leaf tree T_n in terms of the parametrization for two smaller trees.

Proposition 15. *Let T_n be an n -leaf tree. Let $T_n = T_{n-1} * T_3$ be a decomposition of T_n into an $n - 1$ leaf tree and a 3 leaf tree at a cherry. Then*

$$CV(T_n) = CV(T_{n-1}) * CV(T_3).$$

Proof. Consider the parametrization for T_{n-1} written in the usual way as

$$q_{i_1 i_2 \dots i_{n-2} k}^{j_1 j_2 \dots j_{n-2} l} = \sum_{(i_v) \in H} \prod_e e_{i_s(e) i_t(e)}^{j_s(e)}.$$

and the parameterization for the 3 leaf tree T_3

$$r_{k i_{n-1} i_n}^{l j_{n-1} j_n} = \sum_{i_u \in \{0,1\}} \prod_f f_{i_s(f) i_t(f)}^{j_s(f)}$$

where u is the interior vertex of T_3 . Writing the first tensor as Q and the second as R , we have an entry of $P = Q * R$ given by the formula

$$p_{i_1 i_2 \dots i_n}^{j_1 j_2 \dots j_n} = \sum_{k \in \{0,1\}} q_{i_1 i_2 \dots i_{n-2} k}^{j_1 j_2 \dots j_{n-2} l} r_{k i_{n-1} i_n}^{l j_{n-1} j_n}$$

where l satisfies $\sum j_{n-1} + j_n + l = 0 \in \mathbb{Z}_2$. Let \mathbf{e} and \mathbf{f} denote the distinguished edges of T_{n-1} and T_3 respectively which are joined to make the tree T_n . Expanding the expression and regrouping yields

$$\begin{aligned} &= \sum_{k \in \{0,1\}} \left(\sum_{(i_v) \in H} \prod_e e_{i_s(e) i_t(e)}^{j_s(e)} \right) \left(\sum_{i_v \in \{0,1\}} \prod_f f_{i_s(f) i_t(f)}^{j_s(f)} \right) \\ &= \sum_{(i_v) \in H} \sum_{i_u \in \{0,1\}} \sum_{k \in \{0,1\}} \prod_e e_{i_s(e) i_t(e)}^{j_s(e)} \prod_f f_{i_s(f) i_t(f)}^{j_s(f)}. \\ &= \sum_{(i_v) \in H} \sum_{i_u \in \{0,1\}} \prod_{e \neq \mathbf{e}} e_{i_s(e) i_t(e)}^{j_s(e)} \prod_{f \neq \mathbf{f}} f_{i_s(f) i_t(f)}^{j_s(f)} \left(\sum_{k \in \{0,1\}} \mathbf{e}_{i_s(\mathbf{e}) i_k}^l \mathbf{f}_{i_k i_t(\mathbf{f})}^l \right). \end{aligned}$$

The parenthesized expression is the product of the G -matrices \mathbf{e} and \mathbf{f} . Replacing this expression with a new single G -matrix of parameters along the conjoined edge \mathbf{ef} proves that $CV(T_{n-1}) * CV(T_3) \subseteq CV(T_n)$. Now expanding the parameterization given in Lemma 8 as a sum on the vertex u we obtain the other inclusion. □

Now we define a variety ${}^G V(4^{r_1}, 4^{r_2}, 4^{r_3})$ which plays a large role when we extend invariants.

Definition 16. For $l = 1, 2, 3$ let (\mathbf{j}_l) be a string of indices of length r_l . Let ${}_l M$ be an arbitrary G -matrix of size 4^{r_l} where the rows are indexed by $\left\{ \binom{0}{0}, \binom{0}{1}, \binom{1}{0}, \binom{1}{1} \right\}$ and the columns are indexed by the 4^{r_l} indices (\mathbf{j}_l) . Define the parametrization $Q = \psi_{r_1, r_2, r_3}({}_1 M, {}_2 M, {}_3 M)$ by

$$Q_{\mathbf{i}_1 \mathbf{i}_2 \mathbf{i}_3}^{\mathbf{j}_1 \mathbf{j}_2 \mathbf{j}_3} = \sum_{i \in \{0,1\}} {}_1 M_{i \mathbf{i}_1}^{\sigma(\mathbf{j}_1) \mathbf{j}_1} {}_2 M_{i \mathbf{i}_2}^{\sigma(\mathbf{j}_2) \mathbf{j}_2} {}_3 M_{i \mathbf{i}_3}^{\sigma(\mathbf{j}_3) \mathbf{j}_3}$$

if $\sigma(\mathbf{j}_1) + \sigma(\mathbf{j}_2) + \sigma(\mathbf{j}_3) = 0$ and $Q_{\mathbf{i}_1 \mathbf{i}_2 \mathbf{i}_3}^{\mathbf{j}_1 \mathbf{j}_2 \mathbf{j}_3} = 0$ if $\sigma(\mathbf{j}_1) + \sigma(\mathbf{j}_2) + \sigma(\mathbf{j}_3) = 1$. The projective variety that is the Zariski closure of the image of ψ_{r_1, r_2, r_3} is denoted ${}^G V(4^{r_1}, 4^{r_2}, 4^{r_3})$. The affine cone over this variety is $C{}^G V(4^{r_1}, 4^{r_2}, 4^{r_3})$.

Remark 17. By the definition of ${}^G V(4^{r_1}, 4^{r_2}, 4^{r_3})$ any $Q \in {}^G V(4^{r_1}, 4^{r_2}, 4^{r_3})$ is a G -tensor. Furthermore ${}^G V(4, 4, 4)$ is equal to the variety defined by the SSM on the three leaf claw tree $K_{1,3}$.

Besides the fact that ${}^G V(4^{r_1}, 4^{r_2}, 4^{r_3})$ is equal to the SSM when $r_1 = r_2 = r_3 = 1$ the importance of this variety for the strand symmetric model comes from the fact that ${}^G V(4^{r_1}, 4^{r_2}, 4^{r_3})$ contains the SSM for any binary tree as illustrated by the following proposition.

Proposition 18. *Let T be a binary tree and v an interior vertex. Suppose that removing v from T partitions the leaves of T into the three sets $\{1, \dots, r_1\}$, $\{r_1 + 1, \dots, r_1 + r_2\}$, and $\{r_1 + r_2 + 1, \dots, r_1 + r_2 + r_3\}$. Then the SSM on T is a subvariety of ${}^G V(4^{r_1}, 4^{r_2}, 4^{r_3})$.*

In the proposition, the indices in the Fourier coordinates for the SSM are grouped in the natural way according to the tripartition of the leaves.

Proof. In the parametric representation

$$Q_{i_1 i_2 \dots i_n}^{j_1 j_2 \dots j_n} = \sum_{(i_v) \in H} \prod_e e_{i_s(e) i_t(e)}^{j_s(e)}$$

perform the sum associated to the vertex v first. This realizes the G -tensor Q as the sum over the product of entries of three G -tensors. \square

Our goal for the remainder of this section is to prove a result analogous to Theorem 7 in Allman and Rhodes [1]. This theorem will provide a method to explicitly determine the ideal of invariants for ${}^G V(4^{r_1}, 4^{r_2}, 4^{r_3})$ from the ideal of invariants for ${}^G V(4, 4, 4)$. Denote by ${}^G M(2l, 2m)$ the set of $2l \times 2m$ G -matrices. A fundamental observation is that if $r_3' \geq r_3$ then

$$C^G V(4^{r_1}, 4^{r_2}, 4^{r_3'}) = C^G V(4^{r_1}, 4^{r_2}, 4^{r_3}) * {}^G M(4^{r_3}, 4^{r_3'}).$$

Thus, we need to understand the $*$ operation when V and W are “well-behaved” varieties.

Lemma 19. *Let $V \subset {}^G M(2l, 4)$ be a variety and suppose that $V * {}^G M(4, 4) = V$. Let I be the vanishing ideal of V . Let K be the ideal of 3×3 G -minors of the $2l \times 2m$ G -matrix of indeterminates Q . Let Z be $2m \times 4$ G -matrix of indeterminates and*

$$L = \langle \text{coeff}_Z(f(Q * Z)) \mid f \in \text{gens}(I) \rangle.$$

*Then $K + L$ is the vanishing ideal of $W = V * {}^G M(4, 2m)$.*

By a G -minor we mean a minor which involves only the nonzero entries in the G -matrix Q .

Proof. A useful fact is that

$$L = \langle f(Q * A) \mid f \in I, A \in {}^G M(2m, 4) \rangle.$$

Let J be the vanishing ideal of W . By the definition of W , all the polynomials in K must vanish on it. Moreover if $f(Q * A)$ is a polynomial in L , then it vanishes at all the points of the form $P * B$, for any $P \in V$ and $B \in {}^G M(4, 2m)$. Indeed, as $P * B * A \in V$ and $f \in I$ we have $f(P * B * A) = 0$. As all the points of W are of this form, we obtain the inclusion $K + L \subseteq J$. Our goal is to show that $J \subseteq K + L$.

Since $V * {}^G M(4, 4) = V$, we must also have $W * {}^G M(2m, 2m) = W$. This implies that there is an action of $Gl(\mathbb{C}, m) \times Gl(\mathbb{C}, m)$ on W and hence, any graded piece of J , the vanishing ideal of W , is a representation of $Gl(\mathbb{C}, m) \times Gl(\mathbb{C}, m)$. Let J_d be the d -th graded piece of J . Since $Gl(\mathbb{C}, m) \times Gl(\mathbb{C}, m)$ is reductive, we just need to show each irreducible subspace M of J_d belongs to $K + L$. By construction, $K + L$ is also invariant under the action of $Gl(\mathbb{C}, m) \times Gl(\mathbb{C}, m)$ and, hence, it suffices to show that there exists a polynomial $f \in M$ such that $f \in K + L$.

Let $f \in M$ be an arbitrary polynomial in the irreducible representation M . Let P be a $2l \times 4$ G -matrix of indeterminates. Suppose that for all $B \in {}^G M(4, 2m)$, $f(P * B) \equiv 0$. This implies that f vanishes when evaluated at any G -matrix Q which has rank 2 in both components. Hence, $f \in K$.

If $f \notin K$ there exists a $B \in {}^G M(4, 2m)$ such that $f_B(P) := f(P * B) \not\equiv 0$. Renaming the P indeterminates we can take D a matrix in ${}^G(2m, 4)$ formed by ones and zeroes such that $f_B(Q * D) \not\equiv 0$. Since $f \in J$, we must have $f_B(P) \in I$. Therefore $f_B(Q * D) \in L$. Let $B' = D * B \in {}^G M(2m, 2m)$. Although $B' \notin Gl(\mathbb{C}, m) \times Gl(\mathbb{C}, m)$, the representation M must be closed and hence $f(Q * B') = f_B(Q * D) \in M$ which completes the proof. \square

Proposition 20. *Generators for the vanishing ideal of ${}^G V(4^{r_1}, 4^{r_2}, 4^{r_3})$ are explicitly determined by generators for the vanishing ideal of ${}^G V(4, 4, 4)$.*

Proof. Starting with ${}^G V(4, 4, 4)$, apply the preceding lemma three times. Now we will explain how to compute these polynomials explicitly. For $l = 1, 2, 3$ let Z_l be a $4^{r_l} \times 4$ G -matrix of indeterminates. This G -matrix Z_l acts on the $4^{r_1} \times 4^{r_2} \times 4^{r_3}$ tensor Q by G -tensor multiplication in the l -th coordinate. For each $f \in \text{gens}(I)$, where I is the vanishing ideal of ${}^G V(4, 4, 4)$, we construct the polynomials $\text{coeff}_Z f(Q * Z_1 * Z_2 * Z_3)$. That is, we construct the $4 \times 4 \times 4$ G -tensor $Q * Z_1 * Z_2 * Z_3$, plug this into f and expand, and extract, for each Z monomial, the coefficient, which is a polynomial in the entries of Q . Letting f range over all the generators of I determines an ideal L .

We can also flatten the 3-way G -tensor Q to a G -matrix in three different ways. For instance, we can flatten in to a $4^{r_1} \times 4^{r_2+r_3}$ G -matrix grouping the last two coordinates together. Taking the ideal generated by the 3×3 G -minors in these three flattenings yields an ideal K . The ideal $K + L$ generates the vanishing ideal of ${}^G V(4^{r_1}, 4^{r_2}, 4^{r_3})$. \square

5 Extending invariants

In this section we will show how to derive invariants for arbitrary trees from the invariants introduced in section 3. We also introduce the degree 3 determinantal flattening invariants which arise from flattening the n -way G -tensor associated to a tree T under the SSM along an edge of the tree. The idea behind all of our results is to use the embedding of the SSM into the variety ${}^G V(4^{r_1}, 4^{r_2}, 4^{r_3})$.

Let T be a tree with n taxa on the SSM and let v be any interior vertex. Removing v creates a tripartition of the leaves into three sets of cardinalities r_1, r_2 and r_3 , which we may suppose, without loss of generality, are the sets $\{1, \dots, r_1\}, \{r_1 + 1, \dots, r_1 + r_2\}$, and $\{r_1 + r_2 + 1, \dots, r_1 + r_2 + r_3\}$.

Proposition 21. *Let $f(\mathbf{m}_\bullet, \mathbf{n}_\bullet, \mathbf{o}_\bullet, \mathbf{i}_\bullet, \mathbf{j}_\bullet, \mathbf{k}_\bullet)$ be one of the degree 3 invariants for the 3 taxa tree $K_{1,3}$ introduced in Proposition 10. For each $l = 1, 2, 3$ we choose sets of indices $\mathbf{m}_l, \mathbf{i}_l \in \{0, 1\}^{r_1}$, $\mathbf{n}_l, \mathbf{j}_l \in \{0, 1\}^{r_2}$, and $\mathbf{o}_l, \mathbf{k}_l \in \{0, 1\}^{r_3}$ such that $\sigma(\mathbf{m}_l) = m_l$, $\sigma(\mathbf{n}_l) = n_l$ and $\sigma(\mathbf{o}_l) = o_l$. Then $f(\mathbf{m}_\bullet, \mathbf{n}_\bullet, \mathbf{o}_\bullet, \mathbf{i}_\bullet, \mathbf{j}_\bullet, \mathbf{k}_\bullet)$*

$$= \begin{vmatrix} q_{i_1 j_1 k_1}^{m_1 n_1 o_1} & q_{i_2 j_1 k_1}^{m_2 n_1 o_1} & 0 \\ q_{i_1 j_2 k_2}^{m_1 n_2 o_2} & q_{i_2 j_2 k_2}^{m_2 n_2 o_2} & q_{i_3 j_3 k_2}^{m_3 n_3 o_2} \\ q_{i_1 j_2 k_3}^{m_1 n_2 o_3} & q_{i_2 j_2 k_3}^{m_2 n_2 o_3} & q_{i_3 j_3 k_3}^{m_3 n_3 o_3} \end{vmatrix} - \begin{vmatrix} q_{i_1 j_3 k_1}^{m_1 n_3 o_1} & q_{i_2 j_3 k_1}^{m_2 n_3 o_1} & 0 \\ q_{i_1 j_2 k_2}^{m_1 n_2 o_2} & q_{i_2 j_2 k_2}^{m_2 n_2 o_2} & q_{i_3 j_1 k_2}^{m_3 n_1 o_2} \\ q_{i_1 j_2 k_3}^{m_1 n_2 o_3} & q_{i_2 j_2 k_3}^{m_2 n_2 o_3} & q_{i_3 j_1 k_3}^{m_3 n_1 o_3} \end{vmatrix}$$

is a phylogenetic invariant for T .

Proof. The polynomial $f(\mathbf{m}_\bullet, \mathbf{n}_\bullet, \mathbf{o}_\bullet, \mathbf{i}_\bullet, \mathbf{j}_\bullet, \mathbf{k}_\bullet)$ must vanish on the variety ${}^G V(4^{r_1}, 4^{r_2}, 4^{r_3})$. This is because choosing $\mathbf{m}_1, \mathbf{m}_2, \dots$ in the manner specified corresponds to choosing a $3 \times 3 \times 3$ subtensor of Q which belongs to a $4 \times 4 \times 4$ G -subtensor of Q (after flattening to a 3-way tensor). Since ${}^G V(4, 4, 4)$ arises as a projection of ${}^G V(4^{r_1}, 4^{r_2}, 4^{r_3})$ onto this G -subtensor, $f(\mathbf{m}_\bullet, \mathbf{n}_\bullet, \mathbf{o}_\bullet, \mathbf{i}_\bullet, \mathbf{j}_\bullet, \mathbf{k}_\bullet)$ belongs to the corresponding elimination ideal. Since the variety of the SSM for T is contained in the variety ${}^G V(4^{r_1}, 4^{r_2}, 4^{r_3})$, $f(\mathbf{m}_\bullet, \mathbf{n}_\bullet, \mathbf{o}_\bullet, \mathbf{i}_\bullet, \mathbf{j}_\bullet, \mathbf{k}_\bullet)$ is an invariant for the SSM on T . \square

Similarly, we can extend the construction of degree four invariants to arbitrary trees T by replacing the indices in their definition with vectors of indices. We omit the proof which follows the same lines as the preceding proposition.

Proposition 22. *Let $\mathbf{m}, \mathbf{i}_1 \in \{0, 1\}^{r_1}$, $\mathbf{n}_1, \mathbf{j}_1 \in \{0, 1\}^{r_2}$, and $\mathbf{o}_1, \mathbf{k}_1 \in \{0, 1\}^{r_3}$. Then the three 2×2 minors of the following matrix are all degree 4 invariants of the SSM model on the tree T :*

$$\begin{pmatrix} |q_{i_1}^{mn_1 o_1}| & \begin{vmatrix} q_{i_1 j_1}^{mn_1 o_1} \\ q_{i_2 j_2}^{mn_1 o_1} \end{vmatrix} + \begin{vmatrix} q_{i_1 j_2}^{mn_1 o_1} \\ q_{i_2 j_1}^{mn_1 o_1} \end{vmatrix} & |q_{i_2}^{mn_1 o_1}| \\ |q_{i_1}^{mn_2 o_2}| & \begin{vmatrix} q_{i_1 j_3}^{mn_2 o_2} \\ q_{i_2 j_4}^{mn_2 o_2} \end{vmatrix} + \begin{vmatrix} q_{i_1 j_4}^{mn_2 o_2} \\ q_{i_2 j_3}^{mn_2 o_2} \end{vmatrix} & |q_{i_2}^{mn_2 o_2}| \end{pmatrix}.$$

Now we wish to describe the determinantal edge invariants which arise by flattening the G -tensor Q to a matrix along each edge of the tree. As we shall see, their existence is already implied by our previous results, namely Proposition 20. We make the special point of describing them here because they will be useful in the next section.

Let e be an edge in the tree T . Removing this edge partitions the leaves of T into two sets of size r_1 and r_2 . The G -tensor Q flattens to a $4^{r_1} \times 4^{r_2}$ G -matrix R . Denote by \mathcal{F}_e the set of 3×3 G -minors of R .

Proposition 23. *The 3×3 G -minors \mathcal{F}_e are invariants of the SSM on T .*

Proof. The edge e is incident to some interval vertex v of T . These 3×3 G -minors are in the ideal of say ${}^G V(4^{r_1}, 4^{r'_2}, 4^{r'_3})$ associated to flattening the tensor Q to a 3-way G tensor at this vertex. Then by Proposition 18 \mathcal{F}_e are invariants of the SSM on T . \square

6 Reduction to $K_{1,3}$

In this section, we explain how the problem of computing defining invariants for the SSM on a tree T reduces to the problem of computing defining invariants on the claw tree $K_{1,3}$. Our statements and proof are intimately related to the results of Allman and Rhodes [1] and we draw much inspiration from their work.

Given an internal vertex v of T , denote by ${}^G V_v$ the variety ${}^G V(4^{r_1}, 4^{r_2}, 4^{r_3})$ associated to flattening the G -tensor Q to a 3-way tensor according to the tripartition induced by v .

Theorem 24. *Let T be a binary tree. For each $v \in \text{Int}V(T)$ let \mathcal{F}_v be a set of invariants which define the variety ${}^G V_v$ set theoretically. Then*

$$CV(T) = \bigcap_{v \in \text{Int}V(T)} {}^G V_v$$

and hence

$$\mathcal{F}_{\text{flat}}(T) = \bigcup_{v \in \text{Int}V(T)} \mathcal{F}_v$$

are a defining set of invariants for the SSM on T .

The theorem reduces the computation of defining invariants to $K_{1,3}$ since a defining set of invariants for ${}^G V(4^{r_1}, 4^{r_2}, 4^{r_3})$ can be determined from a set of defining invariants for ${}^G V(4, 4, 4) = V(K_{1,3})$. Given the reparametrization

result of Section 4, it will suffice to show the following lemma, about the $*$ operation on G -matrix varieties.

Lemma 25. *Let $V \subseteq {}^G M(2l, 4)$ and $W \subseteq {}^G M(4, 2m)$ be two varieties such that $V = V * {}^G M(4, 4)$ and $W = {}^G M(4, 4) * W$. Then*

$$V * W = (V * {}^G M(4, 2m)) \cap ({}^G M(2l, 4) * W).$$

Proof. Call the variety on the right hand side of the equality U . Since both of the component varieties of U contain $V * W$, we must have $V * W \subseteq U$. Our goal is to show the reverse inclusion. Let $Q \in U$. This matrix can be visualized as a block diagonal matrix:

$$Q = \begin{pmatrix} Q_0 & 0 \\ 0 & Q_1 \end{pmatrix}.$$

Since $Q \in U$ it must be the case that the rank of Q_0 and Q_1 are both less than or equal to 2. Thus we can factorize Q as $Q = R * S$ where $R \in {}^G M(2l, 4)$ and $S \in {}^G M(4, 2m)$. Without loss of generality, we may suppose that the factorization $Q = R * S$ is nondegenerate in the sense that the rank of each of the matrices R and S has only $\text{rank}(Q)$ nonzero rows. Our goal is to show that $R \in V$ and $S \in W$ as this will imply the theorem.

By our assumption that the factorization $Q = R * S$ is nondegenerate, there exists a G -matrix $A \in {}^G M(2m, 4)$ such that $Q * A = R * S * A = R$ (A is called the pseudo-inverse of S). Augmenting the matrix A with extra 0-columns, we get a G -matrix $A' \in {}^G M(2m, 2m)$. Then $Q * A' \in V * {}^G M(4, 2m)$ since Q is and $V * {}^G M(4, 2m)$ is closed under multiplication by G -matrices on the right. On the other hand, the natural projection of $Q * A'$ to ${}^G M(2l, 4)$ is $Q * A = R$. Since the projection $V * {}^G M(4, 2m) \rightarrow {}^G M(2l, 4)$ is the variety V because $V = V * {}^G M(4, 4)$, we have $R \in V$. A similar argument yields $S \in W$ and completes the proof. \square

Now we are in a position to give the proof the Theorem 24.

Proof. We proceed by induction on n the number of leaves of T . If $n = 3$ there is nothing to show since this is the three leaf claw tree $K_{1,3}$. Let T be a binary n taxa tree. The tree T has a cherry T_3 , and thus we can represent the tree $T = T_{n-1} * T_3$ and the resulting variety as $V(T) = V(T_{n-1}) * V(T_3)$ by the reparametrization. Now we apply the induction hypothesis to T_{n-1} and T_3 . The varieties $V(T_{n-1})$ and $V(T_3)$ have the desired representation as

intersections of ${}^G V_v$. By the preceding Lemma, it suffices to show that this representation extends to the variety $V(T_{n-1}) * {}^G M(4, 16)$ and ${}^G M(4^{n-1}, 4) * V(T_3)$. This is almost immediate, since

$${}^G V(4^{r_1}, 4^{r_2}, 4^{r_3}) * {}^G M(4, 4^s) = {}^G V(4^{r_1}, 4^{r_2}, 4^{r_3+s-1})$$

where ${}^G M(4, 4^s)$ acts on a *single index* of ${}^G V(4^{r_1}, 4^{r_2}, 4^{r_3})$ (recall that ${}^G V(4^{r_1}, 4^{r_2}, 4^{r_3})$ can be considered as either a 3-way tensor or an n -way $4 \times \cdots \times 4$ tensor). This equation of varieties applies to each of the component varieties in the intersection representation of $V(T_{n-1})$ and $V(T_3)$ and completes the proof. \square

Acknowledgments

We would like to thank Bernd Sturmfels for some useful conversations about this problem. Marta Casanellas was partially supported by “Ministerio de Ciencia y Tecnología”, BFM2003-06001 and Grant BIO2000-1352-C02-02 of “Plan Nacional I+D” of Spain. Seth Sullivant was supported by a NSF graduate research fellowship. Much of the research in this chapter occurred during a week-long visit to the Universitat de Barcelona, and we are grateful for their hospitality.

References

- [1] Elizabeth S Allman and John A Rhodes, *Phylogenetic ideals and varieties for the general markov model*, math.AG/0401604, 2004.
- [2] S Evans and T Speed, *Invariants of some probability models used in phylogenetic inference*, The Annals of Statistics **21** (1993), 355–377.
- [3] Daniel R. Grayson and Michael E. Stillman, *Macaulay 2, a software system for research in algebraic geometry*, Available at <http://www.math.uiuc.edu/Macaulay2/>, 2002.
- [4] Charles Semple and Mike Steel, *Phylogenetics*, Oxford Lecture Series in Mathematics and its Applications, vol. 24, Oxford University Press, Oxford, 2003. MR MR2060009
- [5] Bernd Sturmfels and Seth Sullivant, *Toric ideals of phylogenetic invariants*, J. Comp. Bio. (2004), to appear.

- [6] L. Szekely, M. Steel, and P. Erdos, *Fourier calculus on evolutionary trees*, Advances in Applied Mathematics **14** (1993), 200–216.