JOURNAL OF ALGEBRAIC STATISTICS
Vol. 6, No. 1, 2015, 17-23
ISSN 1309-3452 – www.jalgstat.com

Tying Up Loose Strands: Defining Equations of the Strand Symmetric Model

Colby Long[1,*], Seth Sullivant[1]

[1] *Department of Mathematics, North Carolina State University, USA*

Abstract. The strand symmetric model is a phylogenetic model designed to reflect the symmetry inherent in the double-stranded structure of DNA. We show that the set of known phylogenetic invariants for the general strand symmetric model of the three leaf claw tree entirely defines the ideal. This knowledge allows one to determine the vanishing ideal of the general strand symmetric model of any trivalent tree. Our proof of the main result is computational. We use the fact that the Zariski closure of the strand symmetric model is the secant variety of a toric variety to compute the dimension of the variety. We then show that the known equations generate a prime ideal of the correct dimension using elimination theory.

2000 Mathematics Subject Classifications: 92D15, 14J99, 60J20

Key Words and Phrases: Algebraic statistics, Phylogenetic invariants, Strand symmetric model

1. Introduction

The strand symmetric model is a phylogenetic model designed to reflect the symmetry inherent in the double-stranded structure of DNA. This symmetry naturally imposes restrictions on the transition probabilities assigned to each edge and imposing only these restrictions gives the general strand symmetric model (SSM). The phylogenetic invariants of a model are algebraic relationships that must be satisfied by the probability distributions arising from the model. Their study was originally proposed as a method for reconstructing phylogenetic trees [4, 10], but they have also been useful theoretical tools in proving identifiability results (see e.g. [2]). Results in [6] imply that to determine generators of the ideal of phylogenetic invariants for any trivalent tree, it suffices to determine them for the claw tree, $K_{1,3}$.

Though the general strand symmetric model itself is not group-based, Casanellas and the second author [3] showed that it is still amenable to the Fourier/Hadamard transform technique of [7, 11]. In the Fourier coordinates, it becomes evident that the parameterization of the model for $K_{1,3}$ is a coordinate projection of the secant variety of the Segre embedding of $\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3$. From this observation, the same authors were able to find 32

[*] Corresponding author.

Email addresses: celong2@ncsu.edu (C. Long), smsulli2@ncsu.edu (S. Sullivant)

degree three and 18 degree four invariants of the homogenous ideal for $K_{1,3}$ and to show that these invariants generate the ideal up to degree four. Whether or not these equations generate the entire ideal was heretofore unknown.

In this paper, we show that these 50 equations in fact generate the entire ideal of the SSM for $K_{1,3}$. First, we use the parameterization of the model after the matrix-valued Fourier transform and the tropical secant dimension technique of Draisma [5] to determine the dimension of the variety of probability distributions arising from the model. Then, using Macaulay2 [9], we show that the ideal generated by these fifty equations defines a variety of the same dimension. Finally, with the aid of symbolic computation we generate a decreasing sequence of elimination ideals demonstrating that the ideal in question is prime. Thus, the variety defined by these equations is irreducible, contains the parameterization, and is of the same dimension as the parameterization, from which the result follows.

2. Phylogenetic Invariants of the SSM model

2.1. Preliminaries

The general strand symmetric model on an n-leaf rooted tree T is a phylogenetic model of 4-state character change. Since the SSM is specifically intended to model DNA evolution, we associate to each node v of the tree a random variable X_v with state space corresponding to the DNA bases {A,C,G,T}. Associated to each edge is a 4×4 transition matrix with rows and columns indexed by the bases. The entry θ_{ij} encodes the probability of changing from character i to j along that edge. In the double helix structure of DNA it is always the case that the bases A and T are paired together and likewise for C and G. So that our model reflects this strand symmetry, we let $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$ be the distribution of the bases at the root, and set $\pi_A = \pi_T$ and $\pi_C = \pi_G$. Additionally, since a character transition in one strand will induce a corresponding transition in the other, we insist

$$\theta_{AA} = \theta_{TT}, \theta_{AC} = \theta_{TG}, \theta_{AT} = \theta_{TA}, \theta_{CA} = \theta_{GT}, \theta_{CC} = \theta_{GG}, \theta_{CG} = \theta_{GC}, \theta_{CT} = \theta_{GA}.$$

The key observation from [3] is that the SSM is a matrix-valued group-based model. Identify the character states of the random variables of a phylogenetic model with elements of $G \times \{0, \dots, l\}$ where G is a finite abelian group and $l \in \mathbb{N}$. Then each character state is indexed by an element $\binom{j}{i}$ where $j \in G$ and $i \in \{0, \dots, l\}$. In these indices, the entries of the transition matrix along edge E are written $E^{j_1 j_2}_{i_1 i_2}$ and the probability that the root is in state $\binom{j}{i}$ is equal to R^j_i.

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

$$E^{j_1 j_2}_{i_1 i_2} = E^{k_1 k_2}_{i_1 i_2}$$

whenever $j_1 - j_2 = k_1 - k_2$ *and the root distribution probabilities satisfy* $R^j_i = R^k_i$.

Let $G = \mathbb{Z}_2$ and $l = 1$, then the following identifications make manifest the matrix-valued group-based structure of the SSM: $A = \binom{0}{0}$, $G = \binom{0}{1}$, $T = \binom{1}{0}$, $C = \binom{1}{1}$.

The tree parameter of an algebraic model determines a polynomial map sending each choice of stochastic parameters into the probability space indexed by n-tuples of the characters. Thus, for the SSM of a tree T, if we let S_T be the space of stochastic parameters we have the following map,

$$\phi_T : S_T \to \Delta^{4^n-1}.$$

If we do not impose the stochastic conditions on the parameters then $\overline{\mathrm{im}(\phi_{\mathrm{T}})}$, where the closure is taken in the Zariski topology, is a variety. In Section 16.1 of [3], the authors detail the group-valued Fourier transform and show how it can be used to obtain a simple parameterization for the closure of the cone over the SSM for $T = K_{1,3}$, denoted $CV(T)$. Letting q_{ijk}^{mno} be the transformed coordinates of the image space, we have

$$\psi : q_{ijk}^{mno} = d_{0i}^{mm} e_{0j}^{nn} f_{0k}^{oo} + d_{1i}^{mm} e_{1j}^{nn} f_{1k}^{oo}$$

if $m+n+o \equiv 0$ in $\mathbb{Z}_2$, and $q_{ijk}^{mno} = 0$ otherwise. Now to determine the defining equations for the SSM for $K_{1,3}$, it is enough to determine the defining equations for $\overline{\mathrm{im}(\psi_{\mathrm{T}})} = CV(T)$. Let I be the ideal generated by the fifty equations found in [3], the rest of the paper will be concerned with proving the following theorem.

Theorem 1. *The vanishing ideal of the strand symmetric model for the graph* $K_{1,3}$ *is minimally generated by* 32 *cubics and* 18 *quartics. The ideal has dimension* 20*, degree* 9024*, and Hilbert series*

$$\frac{1 + 12t + 78t^2 + 332t^3 + 984t^4 + 1908t^5 + 2394t^7 + 1908t^8 + 984t^9 + 332t^{10} + 78t^{11} + 12t^{12} + t^{13}}{(1-t)^{20}}.$$

Note that the Hilbert series suggests that the ideal is Gorenstein though we have not been able to prove this.

2.2. Dimension

A toric variety is a variety that is parametrized by monomials. Let $C \subset CV(T)$ be the toric variety parameterized in each coordinate only by the monomial containing variables with zero in the first entry of the subscript. Thus, $CV(T)$ is the second secant variety of C, denoted $Sec^2(C)$, and we can compute its dimension using existing techniques from [5].

The theorem from [5] which we wish to apply is conveniently formulated for our purposes by Theorem 15 from [1]. We associate to each monomial $x_1^{u_1} x_2^{u_2} \dots x_n^{u_n}$ in the parameterization of a toric variety an integer vector u and let A be the set of these integer vectors. Let $H = \{x \in \mathbb{R}^d : c^T x = e\}$ be a hyperplane in $\mathbb{R}^d$ that splits $\mathbb{R}^d$ into two components which we will label $H^+ = \{x \in \mathbb{R}^d : c^T x > e\}$ and $H^- = \{x \in \mathbb{R}^d : c^T x < e\}$.

In our case, the matrix A is a 12×32 matrix of rank 10, with each column containing exactly threes 1's and nine 0's. If we let $\{e_0^0, e_1^0, e_0^1, e_1^1\}$ denote the standard basis in $\mathbb{R}^{2\times 2}$

then the thirty-two columns of A are

$$\{e_i^m \oplus e_j^n \oplus e_k^o \in \mathbb{R}^{12} : m+n+o \equiv 0 \text{ in } \mathbb{Z}_2\}.$$

For example, the column of A corresponding to the coordinate q_{101}^{110} is given by

$$\begin{pmatrix}0\\0\\0\\1\end{pmatrix} \oplus \begin{pmatrix}0\\0\\1\\0\end{pmatrix} \oplus \begin{pmatrix}0\\1\\0\\0\end{pmatrix}$$

which we write as $(0,0,0,1,0,0,1,0,0,1,0,0)^T$.

Theorem 2. *[1, Theorem 15] Let V_A be a projective toric variety with corresponding set of exponent vectors $A \subset \mathbb{N}^d$. Let H be a hyperplane not intersecting A. Let $A^+ = A \cap H^+$ and $A^- = A \cap H^-$. Then $dim(Sec^2(V_A)) \geq rank(A^+) + rank(A^-) - 1$.*

Recall that I is the ideal generated by the fifty equations found in [3].

Lemma 1. *$dim(CV(T)) = dim(V(I)) = 20$.*

Proof. Regard C as a projective variety so that $C = V_A$ from Theorem 2. The hyperplane defined by the vector $c = (1,0,1,0,1,0,1,0,1,0,1,0)$ and $e = \frac{3}{2}$ gives $|A^+| = |A^-| = 16$ and $\text{rank}(A^+) = \text{rank}(A^-) = 10$. Therefore, by Theorem 2, as a projective variety $\dim(Sec^2(C)) \geq 19$ and as an affine cone $\dim(CV(T)) \geq 20$. Using Macaulay2 we determine that $\dim(V(I)) = 20$, and since $CV(T) \subseteq V(I)$, we must have $\dim(CV(T)) = 20$.

2.3. Primality

In this section we outline our approach for determining if the ideal I is prime. There are algorithms for determining whether or not an ideal is prime implemented in many computer algebra systems. However, these algorithms do not terminate for many of the large ideals confronted in practice, including the ideal I generated by the cubics and quartics contained in $I(CV(K_{1,3}))$. We use the following result from [8] which in certain cases allows one to determine the primality of an ideal by determining the primality of an ideal in fewer variables.

Lemma 2. *[8, Proposition 23] Let k be a field and $J \subset k[x_1, \ldots, x_n]$ be an ideal containing a polynomial $f = gx_1 + h$ with g, h not involving x_1 and g a non-zero divisor modulo J. Let $J_1 = J \cap k[x_2, \ldots, x_n]$ be the elimination ideal. Then J is prime if and only if J_1 is prime.*

Proposition 23 of [8] was stated without proof, so we include a proof of the result for completeness.

Proof. ($\Rightarrow$) It is true in general that the elimination ideal of a prime ideal is prime. Suppose J is prime and let $a, b \in k[x_1, \ldots, x_n] \setminus J_1$ such that $ab \in J_1$. Since $J_1 \subset J$, it must be that either a or b is in $J \setminus J_1$, otherwise it would contradict that J is prime. Therefore, either a or b is in $k[x_1, \dots, x_n] \setminus k[x_2, \dots, x_n]$ and so ab must have some term that involves x_1, which implies $ab \notin J_1$, a contradiction.

($\Leftarrow$) Suppose J_1 is prime but that J is not. Then there must exist $a, b \in k[x_1, \dots, x_n] \setminus J$ with $ab \in J \setminus J_1$. Choose a and b so that ab has minimal x_1-degree among all such pairs. Let d be the x_1-degree of a and d' the x_1-degree of b. Since $ab \in J \setminus J_1$, $d + d' \geq 1$, and so without loss of generality we can assume $d \geq 1$. Write

$$a = h_0 + h_1 x_1 + h_2 x_1^2 + \ldots + h_d x_1^d,$$

where each $h_i \in k[x_2, \dots, x_n]$ and $h_d \neq 0$. Then since $f \in J$ and g is not a zero divisor mod J, $a' := (ga - h_d x_1^{d-1} f)$ is not in J and has x_1-degree strictly less than d. It follows that $a'b$ has x_1-degree strictly less than that of ab. Finally, since ab and f are in J, $a'b = gab - h_d x_1^{d-1} fb$ is in J, contradicting the minimality of the x_1-degree of ab.

Lemma 3. *The ideal I generated by the* 32 *cubics and* 18 *quartics of the general strand symmetric model for $K_{1,3}$ is prime.*

Proof. The proof is obtained by repeated application of Lemma 2. The computations we describe can be found at

`http://www4.ncsu.edu/~smsulli2/Pubs/LooseStrandsWebsite/SSM_Supplement.html`

in the Macaulay2 file `SSM_Supplement` where the symbols 0,1,2, and 3 are substituted for $\binom{1}{1}$, $\binom{1}{0}$, $\binom{0}{1}$, and $\binom{0}{0}$.

First, we let $I_0 = I$. Beginning with $k = 1$, we find a polynomial $f_k = g_k x_k + h_k \in I_{k-1}$, verify that g_k is not a zero-divisor mod I_{k-1}, and then eliminate x_k to obtain the ideal I_k. In this way we generate a decreasing chain of elimination ideals

$$I = I_0 \supset I_1 \supset I_2 \ldots \supset I_{10}.$$

Using the `isPrime` function in Macaulay2, we show that I_{10}, and hence every ideal in the sequence, is prime.

While this is the general outline of our approach, it is actually computationally easier to show that none of the g_k that we encounter is a zero-divisor mod the respective elimination ideal first. Identify the new indices $0, 1, 2$, and 3 with the set of standard basis vectors $\{e_1, e_2, e_3, e_4\}$ and define a multi-grading where the weight of q_{ijk} is $e_{i+1} \oplus e_{j+1} \oplus e_{k+1}$. Let $q_\alpha q_\beta - q_\gamma q_\delta$ be a nontrivial binomial that is homogenous with respect to this grading. For this particular sequence of ideals we are always able to choose $f_k = g_k x_k + h_k$ so that g_k is either such a binomial or a product of such binomials. There are two elementary observations that will be useful:

(i) $g = l_1 l_2$ is a zero-divisor mod J if and only if at least one of l_1 and l_2 is.

(ii) g is not a zero-divisor mod any elimination ideal of J if it is not a zero-divisor mod J.

Thus, to show that none of the g_k is a zero-divisor mod I_{k-1} it is enough to show that none of the homogenous binomials is a zero-divisor mod I.

The symmetry of I enables us to establish this by considering only a small subset of the homogenous binomials. There is a group action of $S_4 \times S_4 \times S_4 \rtimes S_3$ on $Sec^2(Seg(\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3))$, that comes from performing the rank-preserving column and transposition operations. Hence, the same group acts on $I(Sec^2(Seg(\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3)))$, where column operations correspond to changing the indices of the variables and transposition operations correspond to permuting the indices of each variable. Let G be the subgroup of elements of $S_4 \times S_4 \times S_4 \rtimes S_3$ satisfying $g \cdot q_{ijk}^{mno} = q_{i'j'k'}^{m'n'o'}$ with $m + n + o \equiv m' + n' + o'$ in $\mathbb{Z}_2$ for each of the 64 variables. Since

$$I(CV(T)) = I(Sec^2(Seg(\mathbb{P}^3 \times \mathbb{P}^3 \times \mathbb{P}^3))) \cap \mathbb{C}[q_{ijk}^{mno} : m + n + o = 0],$$

G acts on $I(CV(T))$, and since the generators of I generate $I(CV(T))$ up to degree four, G acts on I as well. Let H be the subgroup of G generated by elements that correspond to changing the indices. For example, $h = ((01), (01)(23), (01)) \in H$ interchanges $0 \leftrightarrow 1$ in the first index, $0 \leftrightarrow 1$ and $2 \leftrightarrow 3$ in the second, and $0 \leftrightarrow 1$ in the third so that $h \cdot (q_{021}q_{113} - q_{013}q_{121}) = (q_{130}q_{003} - q_{103}q_{030})$. Then

$$\begin{aligned} H = \langle \quad & ((01), id, id), (id, (01), id), (id, id, (01)), ((23), id, id), (id, (23), id), \\ & (id, id, (23)), ((0213), (0213), id), ((0213), id, (0213)) \rangle \end{aligned}$$

is a 256-element normal subgroup and $G \cong H \rtimes S_3$. One can check that the set of homogeneous binomials partitions into three orbits under the action of G with representatives $q_{002}q_{013} - q_{003}q_{012}$, $q_{002}q_{113} - q_{003}q_{112}$, and $q_{002}q_{120} - q_{020}q_{102}$. In the file `SSM_Supplement` we show that none of the homogeneous binomials is a zero-divisor by showing that none of these three binomials is a zero-divisor mod I.

Having shown that I is prime, we are able to give a short proof of Theorem 1.

Proof. [Proof of Theorem 1] The containment $I \subset I(CV(T))$ implies that $CV(T) \subset V(I)$. By Lemma 3, I is prime and so $V(I)$ is an irreducible variety. By Lemma 1, $CV(T)$ is an irreducible variety contained in an irreducible variety of the same dimension, so $CV(T) = V(I)$ and $I = I(CV(T))$. Knowing explicit generators of the vanishing ideal of the strand symmetric model for the graph $K_{1,3}$, the claims about the rank, degree, and Hilbert Series of the ideal are easily verified by the Macaulay2 code in `SSM_Supplement`.

Acknowledgments

Colby Long was partially supported by the US National Science Foundation (DMS 0954865). Seth Sullivant was partially supported by the David and Lucille Packard Foundation and the US National Science Foundation (DMS 0954865).

References

[1] E.S. Allman, S. Petrovic, J.A. Rhodes, and S. Sullivant. Identifiability of 2-tree mixtures for group-based models. *IEEE/ACM Trans Comput Biol Bioinformatics*, 8(3):710–722, 2011 `http://www.ncbi.nlm.nih.gov/pubmed/20733238`.

[2] E.S. Allman and J.A. Rhodes. The identifiability of tree topology for phylogenetic models, including covarion and mixture models. *J. Comp. Biol.*, 13(5):1101–1113, 2006 `http://www.ncbi.nlm.nih.gov/pubmed/16796553`.

[3] Marta Casanellas and Seth Sullivant. *Algebraic Statistics for Computational Biology*, chapter 16. Cambridge University Press, Cambridge, United Kingdom, 2005.

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

[5] J. Draisma. A tropical approach to secant dimensions. *J. Pure Appl. Algebra*, 212(2):349–363, 2008 `http://www.sciencedirect.com/science/article/pii/S0022404907001429`.

[6] Jan Draisma and Jochen Kuttler. On the ideals of equivariant tree models. *Math. Ann.*, 344(3):619–644, 2009 `http://arxiv.org/abs/0712.3230`.

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

[8] Luis David Garcia, Michael Stillman, and Bernd Sturmfels. Algebraic geometry of bayesian networks. *Journal of Symbolic Computation*, 39(3-4):331–355, March-April 2005 `http://www.sciencedirect.com/science/article/pii/S0747717105000076`.

[9] D.R. Grayson and M.E. Stillman. Macaulay2, a software system for research in algebraic geoemetry. Available at http://www.math.uiuc.edu/Macaulay2/, 2002.

[10] J. A. Lake. A rate-independent technique for analysis of nucleaic acid sequences: evolutionary parsimony. *Molecular Biology and Evolution*, 4:167–191, 1987.

[11] L. Székely, P.L. Erdös, M.A. Steel, and D. Penny. A fourier inversion formula for evolutionary trees. *Applied Mathematics Letters*, 6(2):13–17, 1993 `http://www.sciencedirect.com/science/article/pii/0893965993900047`.