```
gap> tblmod2 = BrauerTable( tbl, 2 );
                      Software for
                            i otbl
                                 Geometry
               terTableRegular(
          BrauerTable(
      gap>
      CharacterTable( "Sym(4)" )
                                 ==> 536
                                  timer=0; // reset timer
(2, \{5, 4\}, 9) => 4
                  Phylogenetic trees
(3, {7, 4}, 1) BANOS, NATHANIEL BUSHEK, RUTH DAVIDSON,
(4, ELEZABETH GROSS, PAMELA E. HARRIS, ROBERT KRONE,
    COLBY LONG, ALLEN STEWART AND ROBERT WALKER
```

```
vol 11
```

## **Phylogenetic trees**

## HECTOR BAÑOS, NATHANIEL BUSHEK, RUTH DAVIDSON, ELIZABETH GROSS, PAMELA E. HARRIS, ROBERT KRONE, COLBY LONG, ALLEN STEWART AND ROBERT WALKER

ABSTRACT: We introduce the package *PhylogeneticTrees* for *Macaulay*2, which allows users to compute phylogenetic invariants for group-based tree models. We provide some background information on phylogenetic algebraic geometry and show how the package *PhylogeneticTrees* can be used to calculate a generating set for a phylogenetic ideal as well as a lower bound for its dimension. Finally, we show how methods within the package can be used to compute a generating set for the join of any two ideals.

MOTIVATION. A central problem in phylogenetics is to describe the evolutionary history of n species from their aligned DNA sequences. One way to do this is through a model-based approach. A phylogenetic model is a statistical model, specified parametrically, of molecular evolution at a single DNA site. We can regard the aligned sequences as a collection of n-tuples of the four DNA bases, one from each site. Each choice of parameters results in a probability distribution on the  $4^n$  possible n-tuples. The goal of model-based reconstruction is to find a choice of parameters that yields a distribution close to the empirical distribution. If we are able to do so then it is reasonable to assume that the model is an accurate reflection of the underlying evolutionary process. Most significantly, we can infer that the underlying tree parameter of the phylogenetic model is the evolutionary tree of the species under consideration.

MATHEMATICAL BACKGROUND. In phylogenetic algebraic geometry, the statistical models under consideration are tree-based Markov models. This means that we assume a Markov process proceeds along a tree with a transition matrix associated to each edge. A  $\kappa$ -state phylogenetic model on an *n*-leaf tree  $\mathcal{T}$ induces a polynomial map from the parameter space  $\Theta_{\mathcal{T}} \subseteq \mathbb{R}^m$  to the probability simplex  $\Delta^{\kappa^n-1}$ ,

$$\psi_{\mathcal{T}}: \Theta_{\mathcal{T}} \to \Delta^{\kappa^n - 1} \subset \mathbb{C}^{\kappa^n}.$$

The image of this map is the set of all probability distributions we obtain by varying the entries of the transition matrices; we refer to the image as the model  $M_T$ . For phylogenetic applications, usually  $\kappa = 2$  or  $\kappa = 4$ .

The Zariski closure of the model  $V_{\mathcal{T}} := \overline{\mathcal{M}_{\mathcal{T}}} \subseteq \mathbb{C}^{\kappa^n}$  is an affine algebraic variety. For the models we consider, the entries of  $\psi_{\mathcal{T}}$  are homogeneous polynomials of uniform degree. Thus,  $V_{\mathcal{T}}$  can be viewed

PhylogeneticTrees version 2.0

MSC2010: primary 13P25; secondary 05C05, 14M25, 92D15.

Keywords: phylogenetic trees, secant ideals, algebraic statistics, toric ideals.

as the affine cone of a projective variety in  $\mathbb{P}^{\kappa^n-1}$ . The ideal  $\mathcal{I}_{\mathcal{T}} := I(V_{\mathcal{T}}) \subseteq \mathbb{C}[x_1, \ldots, x_{\kappa^n}]$  of all polynomials vanishing on  $V_{\mathcal{T}}$  is a homogeneous ideal called the *ideal of phylogenetic invariants*; this ideal carries useful information about the model that can be used for determining model identifiability and performing model selection [E. S. Allman and Sullivant 2011; Allman et al. 2012; Cavender and Felsenstein 1987; Casanellas and Fernández-Sánchez 2006; A. 1987; Long and Sullivant 2015; Matsen et al. 2008; Matsen and Steel 2007; Rhodes and Sullivant 2012; Rusinko and Hipp 2012].

Given two models,  $\mathcal{M}_{\mathcal{T}_1}$  and  $\mathcal{M}_{\mathcal{T}_2}$ , we define the *mixture model*  $\mathcal{M}_{\mathcal{T}_1} * \mathcal{M}_{\mathcal{T}_2}$  to be the image of the map

$$\psi_{\mathcal{T}_1,\mathcal{T}_2} : \Theta_{\mathcal{T}_1} \times \Theta_{\mathcal{T}_2} \times [0,1] \to \Delta^{\kappa^n - 1} \subset \mathbb{C}^{\kappa^n};$$
  

$$\psi_{\mathcal{T}_1,\mathcal{T}_2}(\theta_1, \theta_2, \pi) = \pi \psi_{\mathcal{T}_1}(\theta_1) + (1 - \pi) \psi_{\mathcal{T}_2}(\theta_2).$$
(1)

As before, we take the Zariski closure  $\overline{\mathcal{M}_{\mathcal{T}_1} * \mathcal{M}_{\mathcal{T}_2}} \subseteq \mathbb{C}^{\kappa^n}$ , and now we obtain the algebraic variety  $V_{\mathcal{T}_1 * \mathcal{T}_2} = V_{\mathcal{T}_1} * V_{\mathcal{T}_2}$ , the join of  $V_{\mathcal{T}_1}$  and  $V_{\mathcal{T}_2}$ . The *join* of two algebraic varieties V and W embedded in a common ambient space is the variety

$$V * W := \overline{\{\lambda v + (1 - \lambda)w \mid \lambda \in \mathbb{C}, v \in V, w \in W\}}$$

In the special case when V = W, the join variety V \* V is called the *secant variety of* V. Similarly, given two ideals  $I_1, I_2 \subset \mathbb{C}[x_1, \ldots, x_n]$ , the ideal  $I_1 * I_2 \subset \mathbb{C}[x_1, \ldots, x_n]$  is the *join ideal*. As for varieties, if  $I_1 = I_2$ , the ideal  $I_1 * I_2$  is the *secant ideal*. We refer to [Sturmfels and Sullivant 2006, Section 2] for the definition of  $I_1 * I_2$ , but note the following important property:

$$\mathcal{I}_{\mathcal{T}_1} * \mathcal{I}_{\mathcal{T}_2} = I(V_{\mathcal{T}_1} * V_{\mathcal{T}_2}). \tag{2}$$

Our package provides a means of computing invariants for those working in phylogenetic algebraic geometry. We handle a class of commonly used models called *group-based models* that have special restrictions on the entries of the transition matrices. These entries are indexed by elements of a group and thus are subject to the Fourier–Hadamard coordinate transformation, which makes the parametrization monomial and the ideals toric [Evans and Speed 1993; Székely et al. 1993]. We will refer to the original coordinates, which represent leaf probabilities, as *probability coordinates* and the transformed coordinates as *Fourier coordinates*; furthermore, following the literature, we will use p for probability coordinates and q for Fourier coordinates. For these group-based models, we implement a theoretical construction for inductively determining the ideal of phylogenetic invariants for any tree from the invariants for claw trees [Sturmfels and Sullivant 2005]. We also handle the join and secant ideals formed from these ideals, which allows for computations involving mixture models.

FUNCTIONALITY FOR TORIC PHYLOGENETIC VARIETIES. As an example, let  $\mathcal{T}$  be the four-leaf tree illustrated in Figure 1 and consider the Cavender–Farris–Neyman (CFN) model, a two-state group-based model, on  $\mathcal{T}$ . Then the toric ideal  $\mathcal{I}_{\mathcal{T}}$  is generated in degree 2. Using our package, we can compute a generating set for  $\mathcal{I}_{\mathcal{T}}$  using two different methods, phyloToric42 and phyloToricFP.



Figure 1. Four leaf tree

The first method phyloToric42 calls FourTiTwo.m2, the [Macaulay2] interface to [4ti2], a software package with functionality for computing generating sets of toric ideals. We input  $\mathcal{T}$  by its set of non-trivial splits. In this instance, 01|23 is the only nontrivial split of  $\mathcal{T}$ , which we can enter as either {0, 1} or {2, 3}. The indices on the q's correspond to two-state labelings of the four leaves of  $\mathcal{T}$ :

The second method phyloToricFP computes generators of  $\mathcal{I}_{\mathcal{T}}$  using Theorem 24 in [Sturmfels and Sullivant 2005]; the "FP" in the method name stands for fiber product [Sullivant 2007]. For this example, this theorem allows us to explicitly construct a generating set of  $\mathcal{I}_{\mathcal{T}}$  from generators of  $\mathcal{I}_{K_{1,3}}$ , the ideal associated to the CFN model on the claw tree  $K_{1,3}$ . While our example here is binary, we note that this method is implemented for all trees, binary or not.

The algorithm used by phyloToricFP returns more polynomials than are required to generate the ideal. If we wish to directly compare this ideal to that returned by phyloToric42 we must reconstruct both ideals in the same ring. Thus, we use the function qRing to define the ring of Fourier coordinates and use the option of specifying the ring for our ideals:

```
i7: R = qRing(n,M)
i8: = phyloToric42(n,T,M,QRing=>R) == phyloToricFP(n,T,M,QRing=>R)
o8: = true
```

In our experiments, for most cases, phyloToric42 runs much faster than phyloToricFP. This is likely because we have implemented a naive version of the toric fiber product algorithm from [Sturmfels and Sullivant 2005] with no attempt to avoid producing redundant polynomials. It would be worth investigating if there is a faster implementation of this algorithm. Still, one advantage offered by the fiber product method is the ability to inductively construct a single invariant when computing the entire ideal is infeasible. The method phyloToricRandom returns such a randomly constructed invariant.

The polynomials that are returned by both methods are in Fourier coordinates, however, they can be converted to probability coordinates using the function fourierToProbability. To do so, we must first construct the ring of probability coordinates using pRing. Then the method fourierToProbability returns a ring map that converts polynomials in Fourier coordinates to probability coordinates:

i9: = S = pRing(n,M); i10: = phi = fourierToProbability(S,R,4,M); i11: = f = (vars R)\_(0,0) o11: = q\_(0,0,0,0) i12: = phi(f) o12: = (1/2)\*p\_(0,0,0,0)+(1/2)\*p\_(0,0,0,1)+(1/2)\*p\_(0,0,1,0)+(1/2)\*p\_(0,0,1,1) +(1/2)\*p\_(0,1,0,0)+(1/2)\*p\_(0,1,0,1)+(1/2)\*p\_(0,1,1,0)+(1/2)\*p\_(0,1,1,1) +(1/2)\*p\_(1,0,0,0)+(1/2)\*p\_(1,0,0,1)+(1/2)\*p\_(1,0,1,0)+(1/2)\*p\_(1,0,1,1) +(1/2)\*p\_(1,1,0,0)+(1/2)\*p\_(1,1,0,1)+(1/2)\*p\_(1,1,1,0)+(1/2)\*p\_(1,1,1,1)

FUNCTIONALITY FOR SECANT VARIETIES. Mixtures of group-based phylogenetic models correspond to secants and joins of toric ideals, objects that are of interest in combinatorial commutative algebra, but are notoriously hard to compute. In the methods joinIdeal and secant, we implement the elimination method described in [Sturmfels and Sullivant 2006, Section 2] for computing the join of two homogeneous ideals or the secant of one homogeneous ideal.

Consider now the Jukes-Cantor model on  $\mathcal{T}$  from Figure 1. The phylogenetic ideal for the mixture of  $\mathcal{M}_{\mathcal{T}}$  with itself is the second secant ideal of the homogeneous ideal  $\mathcal{I}_{\mathcal{T}}$ , denoted  $\mathcal{I}_{\mathcal{T}} * \mathcal{I}_{\mathcal{T}}$ . For secants, the method secant takes as input a homogenous ideal and an integer k and returns a generating set for the k-th secant ideal. The method also accepts the optional argument DegreeLimit=>{1}, which computes generators of the ideal only up to degree l. Thus, we can obtain generators of degree 3 or less of  $\mathcal{I}_{\mathcal{T}} * \mathcal{I}_{\mathcal{T}}$  with the following commands. The minimal generating set of SecI3 contains 49 linear invariants; we only print the generators with degree greater than one:

```
i13: = I = phyloToric42(n,T,JCmodel);
i14: SecI3 = secant(I,2,DegreeLimit={3});
i15: toString for i in flatten entries mingens SecI3
list (if (degree i)#0 == 1 then continue; i)
o15 = {q_(0,3,3,0)*q_(3,0,2,1)*q_(3,2,0,1)-q_(0,3,2,1)*q_(3,0,3,0)*q_(3,2,0,1)
+q_(0,3,2,1)*q_(3,0,0,3)*q_(3,2,1,0)-q_(0,3,0,3)*q_(3,0,2,1)*q_(3,2,1,0)
-q_(0,3,3,0)*q_(3,0,0,3)*q_(3,2,3,2)+q_(0,3,0,3)*q_(3,0,3,0)*q_(3,2,3,2)}
```

The degree bound allows for the possibility of obtaining some invariants when computing a generating set for the secant ideal is infeasible. Let  $(\mathcal{I}_{\mathcal{T}} * \mathcal{I}_{\mathcal{T}})_l$  be the ideal generated by the elements of  $\mathcal{I}_{\mathcal{T}} * \mathcal{I}_{\mathcal{T}}$  of degree less than or equal to l. In some instances  $(\mathcal{I}_{\mathcal{T}} * \mathcal{I}_{\mathcal{T}})_l$  may be equal to  $\mathcal{I}_{\mathcal{T}} * \mathcal{I}_{\mathcal{T}}$ . To prove this, we must verify that dim $((\mathcal{I}_{\mathcal{T}} * \mathcal{I}_{\mathcal{T}})_l) = \dim(\mathcal{I}_{\mathcal{T}} * \mathcal{I}_{\mathcal{T}})$  and that  $(\mathcal{I}_{\mathcal{T}} * \mathcal{I}_{\mathcal{T}})_l$  is prime. Assuming we are able to compute  $(\mathcal{I}_{\mathcal{T}} * \mathcal{I}_{\mathcal{T}})_l$ , we can compute its dimension and verify that it is prime. We then know that dim $((\mathcal{I}_{\mathcal{T}} * \mathcal{I}_{\mathcal{T}})_l) \ge \dim(\mathcal{I}_{\mathcal{T}} * \mathcal{I}_{\mathcal{T}})$ , leaving the inequality dim $((\mathcal{I}_{\mathcal{T}} * \mathcal{I}_{\mathcal{T}})_l) \le \dim(\mathcal{I}_{\mathcal{T}} * \mathcal{I}_{\mathcal{T}})$  to show. The method toricSecantDim enables us to do this using a probabilistic method based on Terracini's lemma [1911] to compute a lower bound on dim $(\mathcal{I}_{\mathcal{T}} * \mathcal{I}_{\mathcal{T}})$ .

Using this method, we can show that the secant of the ideal from the previous example is in fact generated in degree less than three:

```
i16: = dim(SecI3)
o16: = 12
i17: = isPrime(SecI3)
o17: = true
i18: = toricSecantDim(phyloToricAMatrix(4,{{0,1}},JCmodel),2))
o18: = 12
```

In the code above, we used phyloToricAMatrix(n,T,JCmodel) to construct the defining integral matrix of the toric ideal. For more details, see the documentation and [Sturmfels 1996]. In this instance, the method outlined is substantially faster than using the secant method without a degree bound.

ADDITIONAL FUNCTIONALITY. Although this package was developed with toric ideals from phylogenetics in mind, the methods secant and joinIdeal can be used for *any* homogeneous ideals. Thus, these can be employed for computations outside of phylogenetic algebraic geometry.

The following models are loaded with the package: the Cavender–Farris–Neyman model, the Jukes– Cantor model, the Kimura 2-parameter model, and the Kimura 3-parameter model. Additionally, some functionality for working with trees is available in this package, which includes the methods edgeCut, vertexCut, edgeContract, internalEdges, internalVertices.

ACKNOWLEDGEMENTS. This work began at the 2016 AMS Mathematics Research Community on "Algebraic Statistics," which was supported by the National Science Foundation under grant number DMS-1321794. Ruth Davidson was supported by NSF DMS-1401591. Elizabeth Gross was supported by NSF DMS-1620109. Robert Walker was primarily supported by an NSF GRF under grant number PGF-031543 and partially supported by the NSF RTG grant 0943832.

SUPPLEMENT. The online supplement contains version 2.0 of PhylogeneticTrees.

## REFERENCES.

[4ti2] 4ti2 team, "4ti2: a software package for algebraic, geometric and combinatorial problems on linear spaces", available at www.4ti2.de.

- [A. 1987] L. J. A., "A rate-independent technique for analysis of nucleic acid sequences: evolutionary parsimony", *Mol. Biol. Evol.* **4**:2 (1987), 167–191.
- [Allman et al. 2012] E. S. Allman, J. A. Rhodes, and S. Sullivant, "When Do Phylogenetic Mixture Models Mimic Other Phylogenetic Models?", *Systematic Biology* **61**:6 (2012), 1049–1059.
- [Casanellas and Fernández-Sánchez 2006] M. Casanellas and J. Fernández-Sánchez, "Performance of a New Invariants Method on Homogeneous and Nonhomogeneous Quartet Trees", *Molecular Biology and Evolution* **24**:1 (10 2006), 288–293.
- [Cavender and Felsenstein 1987] J. A. Cavender and J. Felsenstein, "Invariants of phylogenies in a simple case with discrete states", *Journal of Classification* **4** (1987), 57–71.
- [E. S. Allman and Sullivant 2011] J. A. R. E. S. Allman, S. Petrović and S. Sullivant, "Identifiability of Two-Tree Mixtures for Group-Based Models", *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **8**:3 (2011), 710–722.
- [Evans and Speed 1993] S. N. Evans and T. P. Speed, "Invariants of some probability models used in phylogenetic inference", *Ann. Statist.* **21**:1 (1993), 355–377. MR
- [Long and Sullivant 2015] C. Long and S. Sullivant, "Identifiability of 3-class Jukes–Cantor mixtures", *Adv. in Appl. Math.* **64** (2015), 89–110. MR
- [Macaulay2] D. R. Grayson and M. E. Stillman, "Macaulay2: a software system for research in algebraic geometry", available at http://www.math.uiuc.edu/Macaulay2.
- [Matsen and Steel 2007] F. A. Matsen and M. Steel, "Phylogenetic Mixtures on a Single Tree Can Mimic a Tree of Another Topology", *Systematic Biology* **56**:5 (10 2007), 767–775.
- [Matsen et al. 2008] F. A. Matsen, E. Mossel, and M. Steel, "Mixed-up trees: the structure of phylogenetic mixtures", *Bull. Math. Biol.* **70**:4 (2008), 1115–1139. MR
- [Rhodes and Sullivant 2012] J. A. Rhodes and S. Sullivant, "Identifiability of large phylogenetic mixture models", *Bull. Math. Biol.* **74**:1 (2012), 212–231. MR
- [Rusinko and Hipp 2012] J. P. Rusinko and B. Hipp, "Invariant based quartet puzzling", *Algorithms for Molecular Biology* 7:1 (2012).
- [Sturmfels 1996] B. Sturmfels, *Gröbner bases and convex polytopes*, University Lecture Series **8**, American Mathematical Society, Providence, RI, 1996. MR
- [Sturmfels and Sullivant 2005] B. Sturmfels and S. Sullivant, "Toric ideals of phylogenetic invariants", *Journal of Computational Biology* **12**:2 (2005), 204–228.
- [Sturmfels and Sullivant 2006] B. Sturmfels and S. Sullivant, "Combinatorial secant varieties", *Pure Appl. Math. Q.* 2:3, Special Issue: In honor of Robert D. MacPherson. Part 1 (2006), 867–891. MR
- [Sullivant 2007] S. Sullivant, "Toric fiber products", J. Algebra 316:2 (2007), 560-577. MR
- [Székely et al. 1993] L. A. Székely, P. L. Erdős, M. A. Steel, and D. Penny, "A Fourier inversion formula for evolutionary trees", *Appl. Math. Lett.* **6**:2 (1993), 13–16. MR
- [Terracini 1911] A. Terracini, "Sulle  $V_k$  per cui la varieta degli  $S_h(h+1)$ -seganti ha dimensione minore dell'ordinario", *Rend. Circ. Mat. Palermo* (1911).

| RECEIVED: 16 Nov 2016 | REVISED: 4 Jan 2020 | ACCEPTED: 8 Aug 2020 |
|-----------------------|---------------------|----------------------|
|                       |                     |                      |

HECTOR BAÑOS:

6

hbanos@gatech.edu

School of Mathematics, Georgia Institute of Technology, Atlanta, GA, United States

NATHANIEL BUSHEK:

nbushek@css.edu

Department of Mathematics and Physics, The College of St. Scholastica, Duluth, MN, United States

**RUTH DAVIDSON:** 

## redavid2@illinois.edu

Department of Mathematics, University of Illinois Urbana-Champaign, Urbana, IL, United States

ELIZABETH GROSS:

egross@hawaii.edu

Department of Mathematics, University of Hawaii'i at Manoa, Honolulu, HI, United States

PAMELA E. HARRIS:

pamela.e.harris@williams.edu Department of Mathematics and Statistics, Williams College, Williamstown, MA, United States

ROBERT KRONE:

rckrone@ucdavis.edu Department of Mathematics, UC Davis, Davis, CA, United States

COLBY LONG:

clong@wooster.edu Department of Mathematics and Computational Sciences, The College of Wooster, 308 E. University Street, Wooster, OH 44691, United States

ALLEN STEWART:

stewaral@seattleu.edu Department of Mathematics, Seattle University, Seattle, WA, United States

**ROBERT WALKER:** 

rwalker@math.wisc.edu Department of Mathematics, University of Wisconsin–Madison, Madison, WI, United States



JOURNAL OF SOFTWARE FOR ALGEBRA AND GEOMETRY vol 11, no 1, 2021

| Phylogenetic trees<br>Hector Baños, Nathaniel Bushek, Ruth Davidson, Elizabeth Gross, Pamela E.<br>Harris, Robert Krone, Colby Long, Allen Stewart and Robert Walker   |     |
|--|-----|
| Software for doing computations in graded Lie algebras<br>Clas Löfwall and Samuel Lundqvist  | 9   |
| The relative canonical resolution: Macaulay2-package, experiments and conjectures<br>Christian Bopp and Michael Hoff   | 15  |
| <i>The FrobeniusThresholds package for Macaulay2</i><br>Daniel J. Hernández, Karl Schwede, Pedro Teixeira and Emily E. Witt  | 25  |
| Computing theta functions with Julia<br>Daniele Agostini and Lynn Chua   | 41  |
| Decomposable sparse polynomial systems<br>Taylor Brysiewicz, Jose Israel Rodriguez, Frank Sottile and Thomas Yahl  | 53  |
| A package for computations with sparse resultants<br>Giovanni Staglianò  | 61  |
| <i>ExteriorModules: a package for computing monomial modules over an exterior algebra</i><br>Luca Amata and Marilena Crupi   | 71  |
| The Schur–Veronese package in Macaulay2<br>Juliette Bruce, Daniel Erman, Steve Goldstein and Jay Yang  | 83  |
| admcycles - a Sage package for calculations in the tautological ring of the moduli space of stable curves  | 89  |
| Vincent Delecroix, Johannes Schmitt and Jason van Zelm   |     |
| Coding theory package for Macaulay2<br>Taylor Ball, Eduardo Camps, Henry Chimal-Dzul, Delio Jaramillo-Velez, Hiram<br>López, Nathan Nichols, Matthew Perkins, Ivan Soprunov, German Vera-Martínez<br>and Gwyn Whieldon | 113 |
| Threaded Gröbner bases: a Macaulay2 package<br>Sonja Petrović and Shahrzad Zelenberg   | 123 |
| Standard pairs of monomial ideals over nonnormal affine semigroups in SageMath<br>Byeongsu Yu  | 129 |
| Computations with rational maps between multi-projective varieties<br>Giovanni Staglianò   | 143 |