PHYLOGENETIC INVARIANTS FOR GROUP-BASED MODELS

MARIA DONTEN-BURY AND MATEUSZ MICHAŁEK

Abstract. In this paper we investigate properties of algebraic varieties representing group-based phylogenetic models. We give the (first) example of a nonnormal general group-based model for an abelian group. Following Kaie Kubjas [Kub10] we also determine some invariants of group-based models showing that the associated varieties do not have to be deformation equivalent. We propose a method of generating many phylogenetic invariants and in particular we show that our approach gives the whole ideal of the claw tree for 3-Kimura model under the assumption of the conjecture of Sturmfels and Sullivant [SS05]. This, combined with the results in [SS05], would enable to determine all phylogenetic invariants for any tree for 3-Kimura model and possibly for other group-based models.

1. Introduction

Phylogenetics is a science that tries to reconstruct the history of evolution. It is strongly connected with many branches of mathematics including algebraic geometry. To each possible history of evolution, represented by a tree, one can associate an algebraic variety, whose points correspond to possible probability distributions on the DNA states of the living species. For a detailed introduction the reader is advised to look in [PS05] and for an algebraic point of view in [ERSS04].

Biologists are mostly interested in phylogenetic invariants, that are polynomials defining the variety. It is very hard to find them in general, however for some special models of evolution much progress has been made. In this paper we are dealing with a large class of equivariant models [DK09] – so called G-models, that are represented by toric varieties (see [Mic10]). The most influential paper in this area is [SS05], where the authors gave the description of the generators of the ideal, assuming that the ideal of the claw tree is known. Unfortunately not much is known on the ideals of the claw trees. In particular we do not even know if the degree in which they are generated is bounded

The second author is supported by a grant of Polish MNiSzW (N N201 413 539).

while the number of leaves grows to infinity (Conjecture 1 and 2 in [SS05]). In this paper we propose a method of finding the ideals of the claw trees using a geometric approach. We conjecture that the varieties associated to large claw trees are intersections of varieties

associated to trees of smaller valency. This would enable to recursively generate ideals. An interesting fact is that we can show that our conjecture is equivalent to the one made by Sturmfels and Sullivant for the 3-Kimura model (for the details see section 4).

We also try to investigate properties of algebraic varieties representing phylogenetic models. In particular we give an example of a model associated to an abelian group that gives a non-normal variety. The results on normality are strongly connected to deformation problems. It is well-known that algebraic varieties representing trivalent trees with the same number of leaves are deformation equivalent for the binary model. The original geometric proof can be found in [BW07] and a new, more combinatorial one, in [Ilt10]. A new result of Kaie Kubjas shows that this is not true for the 3-Kimura model [Kub10]. The idea of the proof is to calculate the number of integer points in nP, where P is the polytope associated to the algebraic model of phylogenetic tree. This task is much easier for normal varieties (in this case we obtain Hilbert-Ehrhart polynomial of the algebraic model).

One of the tools that we use is a program that computes the polytope defining a toric variety for a given tree and a group (see section 2.1). Our program, implementing the algorithm described in [Mic10], can be found at `http://www.mimuw.edu.pl/~marysia/polytopes` (with a detailed instruction and specification of the input and output data format).

Acknowledgements

The authors would like to thank very much Jarosław Wiśniewski for introducing them to the topic and encouraging them to study it further. We would also like to thank Laurent Manivel for useful remarks.

2. G-models

The first idea of G-models appears in [BDW09] and precise definitions can be found in [Mic10]. A G-model is an algebraic variety associated to a tree and a group G with a normal, abelian subgroup H. We assume that the tree is rooted and the edges are directed away from the root. In phylogenetics the tree describes the history of evolution. The groups encode possible mutation mechanism. We assume that G acts on the set of states A and that by this action the subgroup H acts transitively and freely. The case when $G = H$ is classical and was a subject of many studies – see [SS05] and references therein.

In the general case the G-model is known to be toric (but not necessarily normal) and the vertices of the defining polytope correspond to networks [Mic10].

Definition 2.1. *Let O be the set of orbits of the action of G on the dual group H^*. A network is a function $n : E \to O$, where E is the set of edges of the tree. Moreover we require that for any inner vertex v of the tree, e_0 an incoming edge and $e_1, \dots, e_k$ outcoming edges, there exist characters $\chi_i \in n(e_i)$ such that*

$$\chi_0 = \chi_1 + \cdots + \chi_k.$$

We say that the signed sum of characters around each inner vertex is the trivial character.

Let us note that given a network n we can choose representatives $\chi_e \in n(e)$ for each edge, such that the signed sum of characters χ_e around each inner vertex is trivial. We see also that if $G = H$ then the set of orbits is the same as H^*. Moreover in this case to define a network it is enough to define it on all but one leaves, and then expand it using inner vertices. In this case the sum of characters associated to leaves is trivial.

2.1. **Algorithm of finding the polytope.** In [Mic10, Sect. 4] there is a description of a simple algorithm which for a given G-model computes the coordinates of the vertices of the polytope related to this model. For all our computations we need to pass from an abstract model to a polytope as a first step. Hence we start describing the computational results from a few remarks about the implementation of this algorithm for G abelian.

There are two non obvious points in the implementation. One is step 2 of the algorithm: making a choice of an outcoming edge from each vertex (the tree is rooted and the edges are directed from the root). It is much easier to choose incoming edge for each vertex except the root, as this choice is almost canonical (depends only on the rooting), so does not have to be stored in the memory. The second interesting point is how to write the program to obtain the complexity $O(|N||G|^{|E \setminus N|})$ (E is the set of edges and N is the set of inner vertices of the tree), as predicted in [Mic10]. This requires group operations being performed in unit time. We can easily achieve this by precomputing group operations and storing the table of results – groups we want to work with are small enough. At the moment the program can operate only on a few groups defined in the source, but it is not very complicated to make it possible to read the permutation group from the file (and precompute group operations).

As a result we have a fast program which takes a tree in a simple text format as an input and allows to choose one of the groups from the library. It computes the list of vertices of a polytope associated to the input model and writes it to a file. It also enables the user to work with this polytope, given as an object of an inner class, in the further computations.

However the current version is correct only for abelian groups, there is also a procedure of determining the polytope for a general G-model. An example is given in [Mic10] (and in general the algorithm works in a very similar way). We will probably implement the extended version in the future. By now we needed it only in a few cases, simple enough to proceed without a program.

3. Computational results

3.1. **Example of non-normal G-model.** Knowing that the projective variety associated to a G-model is toric, it is natural to ask whether it is normal. Computations described in [Mic10] have shown that G-models for the groups $\mathbb{Z}_2$, $\mathbb{Z}_3$, $\mathbb{Z}_4$ and $\mathbb{Z}_2 \times \mathbb{Z}_2$ are normal, but 2-Kimura model is not normal. We are interested in the question whether all models for abelian (or at least cyclic) groups are normal. Now, using the previously described program and Polymake (see [GJ00]) we are able to check normality for a few more models.

More precisely, because of Lemma 5.1 from [Mic10], it is enough to check normality for the tripod (a tree with one inner vertex and three leaves). Hence if we check normality for the chosen group and the tripod, we know whether all algebraic varieties for this group and any trivalent tree are normal. Using our program we can obtain the set of vertices of the polytope related to the investigated group and the tripod. Finally we apply Polymake [GJ00] to check the normality of this polytope (in the lattice generated by its vertices). Thus we obtain

Computation 3.1. *The polytope associated with G-model for the tripod and the group $G = H = \mathbb{Z}_6$ is not normal. Hence the algebraic variety representing this model is not normal.*

In particular, the class of abelian models contains non-normal models. We believe it can be difficult to characterize the class of groups for which G-models are normal, or even to determine a big (infinite) class of normal, toric G-models. On the other hand one has the following result:

Proposition 3.2. *Let T be a phylogenetic tree and let G_1 be a subgroup of an abelian group G_2. If the variety corresponding to the tree T and group G_1 is not normal then the variety correpsonding to the tree T and group G_2 is also not normal.*

Proof. Let M_i be a lattice whose basis is indexed by pairs of an edge of a tree and an element of the group G_i. The inclusion $G_1 \subseteq G_2$ gives us a natural injective morphism $f : M_1 \to M_2$. Let $P_i \subset M_i$ be the polytope associated to the model for the tree T and group G_i. Let $\tilde{M}_i \subset M_i$ be a sublattice spanned by vertices of the polytope P_i.

As P_1 is not normal in the lattice spanned by its vertices, there exists a point $x \in nP_1 \cap \tilde{M}_1$, that is not a sum of n vertices of the polytope P_1. Let us consider $y = f(x)$. The vertices of P_1 are mapped to vertices of P_2. We see that $y \in nP_2 \cap \tilde{M}_2$. If P_2 was normal in $\tilde{M}_2$ we would be able to write $y = \sum_{i=1}^{n} q_i$ with $q_i \in P_2$.

Let us notice that each point in the image $f(M_1)$ has got zero on each entry of the coordinates indexed by any edge and any element of the group $g \in G_2 \setminus G_1$. In particular y has got zero on these entries. As all entries of all vertices of P_2 are nonnegative, this proves that all entries indexed by any edge and any element of the group $g \in G_2 \setminus G_1$ are zero for q_i. However, we see that vertices of P_2 that have got all non zero entries on coordinates indexed by pairs of an edge and an element $g \in G_1$ are in the image of P_1. Hence $q_i = f(p_i)$ for $p_i \in P_1$. We see that $x = \sum p_i$, which is impossible. □

In particular we see that all abelian groups G such that $|G|$ is divisible by 6 give rise to non-normal models.

3.2. **Hilbert-Ehrhart polynomials.** The binary model (for trivalent trees) has an interesting property, stated and proved in [BW07]: an elementary mutation of a tree gives a deformation of the associated varieties (see Construction 3.23). This implies that binary models of trivalent trees with the same number of leaves are deformation equivalent (Theorem 3.26 in [BW07]). As it was not obvious what to expect for other G-models, we computed Hilbert-Ehrhart polynomials, which are invariants of deformation, in some simple cases.

3.2.1. *Numerical results.* We checked models for two different trees with six leaves (this is the least number of leaves for which there are non-isomorphic trees, exactly two), the *snowflake* and the *3-caterpillar*. The most interesting ones were the cases of the biologically meaningful 2-Kimura and 3-Kimura models.

The value of the Ehrhart polynomial of a polytope P for a natural number n is the number of lattice points in nP. Thus one way to determine the Hilbert-Ehrhart polynomial of a G-model is to compute numbers of lattice points in some multiples of its polytope. Even if it is not possible to get enough data to determine the polynomials (eg. because numbers are too big), sometimes we can say that polynomials for two models are not equal, because their values for some n are different.

Before we completed our computations, Kaie Kubjas computed numbers of lattice points in the third dilations of the polytopes for 3-Kimura model on the *snowflake* and the *3-caterpillar* with 6 leaves and got 69248000 and 69324800 points respectively ([Kub10]). Thus she proved that varieties associated with these models are not deformation equivalent.

Our computations confirm her results as for the 3-Kimura model and also give the following

Computation 3.3. *The varieties associated with 2-Kimura models for the snowflake and the 3-caterpillar trees have different Ehrhart polynomials. In the second dilations of the polytopes there are 56992 lattice points for the snowflake and 57024 for the 3-caterpillar.*

Also the pairs of varieties associated with G-models for the snowflake and the 3-caterpillar trees and

(1) $G = H = \mathbb{Z}_3$,
(2) $G = H = \mathbb{Z}_4$,
(3) $G = H = \mathbb{Z}_5$,
(4) $G = H = \mathbb{Z}_7$

have different Hilbert-Ehrhart polynomials and therefore are not deformation equivalent. (For these pairs G-models are normal, which can be checked using Polymake.) The precise results of the computations are presented in the Appendix.

In the cases of

(1) $G = H = \mathbb{Z}_8$,
(2) $G = H = \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$,
(3) $G = H = \mathbb{Z}_9$

the varieties have got different Hilbert functions. We were not able to check if they are normal, however if they are then the Hilbert-Ehrhart polynomials are different.

3.2.2. *Some technical details.* We tried two methods of computing numbers of lattice points in dilations of a polytope. The first attempt was the direct method: constructing the list of lattice points in nP by adding vertices of P to lattice points in $(n-1)P$ and reducing repeated entries. This method is not efficient, it needs a lot of memory to work. At first it did not work good enough to obtain numbers of lattice points in $3P$ for 3-Kimura, but after a few technical upgrades (encoding sequences of coordinates as decimal numbers or, better, numbers in the system with the base 4) we were able to confirm Kaie Kubjas' results. The main problem with this method, apart from its inefficiency, is the fact that it works correctly only for normal polytopes. This follows just from the definition of normality for polytopes. As we wanted to investigate 2-Kimura model, we had to implement another algorithm.

The second idea is to compute inductively the relative Hilbert polynomials, i.e. number of points in the n dilation of the polytope intersected with the fiber of the projection onto the group of coordinates that correspond to a given leaf. Our approach is quite similar to the methods used in [Kub10] and [Sul07].

First we compute two functions for the tripod. Let $P \subset \mathbb{Z}^{3m} \cong \mathbb{Z}^m \times \mathbb{Z}^m \times \mathbb{Z}^m$ be the polytope associated to a tripod. Let $pr_i : \mathbb{Z}^{3m} \cong \mathbb{Z}^m \times \mathbb{Z}^m \times \mathbb{Z}^m \to \mathbb{Z}^m$ be a projection onto the i-th group of coordinates. We distinguish one edge of the tripod corresponding to the third group of coordinates in the lattice. Let f be a function such that $f(a)$ for $a = (a_1, \dots, a_m) \in \mathbb{Z}^m$ is the number of lattice points in $(a_1 + \dots + a_m)P$ that project to a by pr_3. We compute $f(a)$ for sufficiently many a to proceed with the algorithm.

Example 3.4. *The polytope P for the binary model has the following vertices:*

$$\begin{aligned} v_1 &= (0,1,0,1,0,1), \\ v_2 &= (0,1,1,0,1,0), \\ v_3 &= (1,0,0,1,1,0), \\ v_4 &= (1,0,1,0,0,1). \end{aligned}$$

These are only integral points in P. In this case $f(1,0) = 2$ because there are only two points, $(1,0,0,1,1,0)$ and $(0,1,1,0,1,0)$, that are in $1P = P$ and project to $(1,0)$ via the third projection.

The function f will be our base for induction. Next, we need to know how many points are there in the fiber of the projection onto two distinguished leaves. Let g be a function such that $g(a,b)$ for $(a,b) = (a_1, \dots, a_m, b_1, \dots, b_m) \in \mathbb{Z}^m \times \mathbb{Z}^m$ is the number of lattice points in $(a_1 + \dots + a_m)P$ that project to a by pr_3 and to b by pr_2. We compute $g(a,b)$ for sufficiently many (a,b) to proceed with the algorithm.

Let T be a tree with a corresponding polytope P and a distinguished leaf l. Let h be a function such that $h(a)$ for $a = (a_1, \ldots, a_m) \in \mathbb{Z}^m$ is equal to the number of points in the fiber of the projection corresponding to leaf l of $(a_1 + \cdots + a_m)P$ onto a. We construct a new tree T' by attaching a tripod to a chosen leaf of T. We call T' a join of T and the tripod. The chosen leaf of T' will be one of the leaves of the attached tripod. As proved in [BW07], [SS05], [Mic10], [Sul07] (depending on the model), the polytope associated to a join of two trees is a fiber product of the polytopes associated to these trees. Thus we can calculate the function h' for T' by a following rule: $h'(a) = \sum_b g(a,b)h(b)$, where the sum is taken over all $b \in \mathbb{Z}^m$ such that $g(a,b) \neq 0$.

This allows us to compute inductively the relative Hilbert polynomial. The last tripod could be attached in the same way. Then one obtains the Hilbert function from relative Hilbert functions simply by summing up over all possible projections. However, it is better to do the last step in a different way.

Suppose that as before we are given a tree T with a distinguished leaf l and a corresponding relative Hilbert function h. We compute the Hilbert function of the tree T' that is a join of the tree T and a tripod using the equality $h'(n) = \sum_a f(a)h(a)$, where $a = (a_1, \ldots, a_m)$ and $\sum a_i = n$. The function f is the basis for induction introduced above.

Thus, decomposing the *snowflake* and the *3-caterpillar* trees to joins of tripods, we can inductively compute (a few small values of) the corresponding Hilbert functions. This method works also for non-normal models, if only the Hilbert function for the tripod can be computed. In particular, for 2-Kimura model the computations turned out to be possible, because its polytope for the tripod is quite well understood (see [Mic10], 5.4), at least to describe fully its second dilation. This way we obtained the results of 3.3.

4. Phylogenetic invariants

In this section we investigate the most important objects of phylogenetic algebraic geometry – ideals of phylogenetic invariants. The main problem in this area is to give an effective description of the whole ideal of the variety associated to a given model on a tree. Our task is to find an efficient way to compute generators of these ideals.

We suggest a way of obtaining all phylogenetic invariants of a claw tree of a G-model - more precisely we conjecture that our invariants generate the whole ideal of the variety. These, together with the results of [SS05] could provide an algorithm listing all generators of the ideal of phylogenetic invariants for any tree and for any G-model (so in particular for a general group-based model).

4.1. **Inspirations.**

The inspirations for our method were the conjectures made by Sturmfels and Sullivant in [SS05]. They are still open but, as we will see, they strongly support our ideas. In particular, we will prove later that our algorithm works for the 3-Kimura model if we assume that the weaker conjecture made in [SS05] holds.

First we introduce some notation. Let $K_{n,1}$ be a claw tree with n leaves. Let $\phi(G,n) = d$ be the least natural number such that the ideal associated to $K_{n,1}$ for the group based model G is generated in degree d. The phylogenetic complexity of the group G is defined as $\phi(G) = sup_n \phi(G,n)$. Based on some numerical results Sturmfels and Sullivant suggested the following conjecture:

Conjecture 4.1. *For any abelian group G we have $\phi(G) \leq |G|$.*

This conjecture was separately stated for the 3-Kimura model, that is for $G = \mathbb{Z}_2 \times \mathbb{Z}_2$.

Still very little is known about the function ϕ apart from the case of the binary model (see also [CP07]):

Proposition 4.2 (Sturmfels, Sullivant)**.** *In case of the binary model $\phi(Z_2) = 2$.*

There are also some computational results – to the table in [SS05] presenting the computations presenting the computations made by Sturmfels and Sullivant a few cases can be added.

Computation 4.3. *Using 4ti2 software* [tt] *we obtained the following:*

- $\phi(6, \mathbb{Z}_3) = 3$,
- $\phi(4, \mathbb{Z}_5) = 4$,
- $\phi(3, \mathbb{Z}_8) = 8$,
- $\phi(3, \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2) = 8$.

For the 3-Kimura model we do not even know whether the function ϕ is bounded. As we will see later, this conjecture is strongly related to the one stated in the next section.

4.2. **A method for obtaining phylogenetic invariants.**

We propose a method that is inspired by the geometry of the varieties we consider. First we have to introduce some notation.

Definition 4.4. *Let V_i be the set of vertices of a tree T_i for $i = 1, 2$. Let e be an inner edge of T_2 joining $v_1, v_2 \in V_2$. We say that the tree T_1 is obtained from the tree T_2 by* contraction of an edge e *if there is a vertex $v \in V_1$ such that*

- $V_1 = \{v\} \cup (V_2 \setminus \{v_1, v_2\})$,
- *v is connected to all the vertices to which v_1 or v_2 are connected,*
- *the edges between other vertices are the same in both trees.*

In such a situation we say that T_2 is a prolongation of T_1.

Remark 4.5. *Note that these definitions are not the same as the definitions of flattenings introduced in* [AR08] *and further studied in* [DK09].

Let us see that in this setting the variety $X(T_1)$ associated to the tree T_1 is in a natural way a subvariety of $X(T_2)$. Notice that we can identify sockets of both varieties, as we may identify their leaves, so both varieties are contained in $\mathbb{P}^s$, where s is the number of sockets. The natural inclusion corresponds to the projection of character lattices: we forget all the coordinates corresponding to the edge joining the vertices v_1 and v_2. Now the following conjecture is natural:

Conjecture 4.6. *The variety $X(K_{n,1})$ is equal to the (scheme theoretic) intersection of all the varieties $X(T_i)$, where T_i is a prolongation of $K_{n,1}$ that has only two inner vertices, both of them of valency at least three.*

As $X(K_{n,1})$ is a subvariety of $X(T_i)$ for any prolongation T_i one inclusion is obvious. Note also that the valency condition is made, because otherwise the conjecture would be obvious – one of the varieties that we intersect would be equal to $X(K_{n,1})$ (contraction of a vertex of degree 2 does not change the corresponding variety). All T_i have got a strictly smaller maximal valency than $K_{n,1}$, so if the conjecture holds then we can inductively use Theorem 23 of Sturmfels and Sullivant [SS05] (see also Theorem 12 [Sul07]) to obtain all phylogenetic invariants for a given model for any tree of any valency, knowing just the ideal of the tripod. More precisely, if 4.6 holds then the degree in which the ideals of claw trees are generated cannot grow when the number of leaves gets bigger. In such a case the ideal of $X(K_{n,1})$ is just the sum of ideals of trees with smaller valency. This means that $\phi(G) = \phi(G, 3)$ which can be computed in many cases. In particular, the conjecture 4.6 implies all cases of the conjecture 4.1 in which we can compute $\phi(G, 3)$ - this includes the most interesting 3-Kimura model.

Of course one may argue that the conjecture 4.6 above is too strong to be true. Later we will prove it for the binary model. We will also consider two modifications of this conjecture to weaker conjectures that can still have a lot of applications. The first modification just states that the conjecture 4.6 holds for n large enough.

Proposition 4.7. *The conjecture 4.6 holds for n large enough if and only if the function ϕ is bounded.*

Proof. One implication is obvious. Suppose that 4.6 holds for $n > n_0$. We choose such d that the ideals associated to $K_{1,l}$ are generated in degree m for $l \leq n_0$. Using 4.6 and the results of [SS05] we can describe the ideal associated to K_n as the sum of ideals generated in degree m. It follows that this ideal is also generated in degree m, so the function ϕ is bounded by m.

For the other implication let us assume that $\phi(n) \leq m$. Let us consider any binomial B that is in the ideal of the claw tree and is of degree less or equal to m. We prove that B belongs to the ideal of some prolongation of a tree T, which is in fact more than the statement of conjecture 4.6.

Such a binomial can be described as a linear relation between (at most m) vertices of the polytope of this variety. Each vertex is given by an association of orbits of characters to edges such that there exist representatives of orbits that sum up to a trivial character. Let us fix such representatives, so that each vertex is given by n characters summing up to a trivial character.

Now the binomial B can be presented as a pair of matrices A_1 and A_2 with characters as entries. Each column of the matrices is a vertex of the polytope. The matrices have got at most m columns and exactly n rows. Let us consider the matrix $A = A_1 - A_2$, that is entries of the matrix A are characters that are differences of entries of A_1 and A_2. We can subdivide the first column of A into groups of at most $|H|$ elements summing up to a trivial character. Then inductively we can subdivide the rows into groups of at most $|H|^i$ elements summing up to a trivial character in each column up to the i-th one.

For $n > |H|^m + 1$ we can find a set S of rows of A such that the characters sum up to a trivial character in each column restricted to S, such that both the cardinality of S and of its complement are greater then 1. Note that the sums of the entries lying in a chosen column and in the rows in S are the same in A_1 and A_2. Therefore, adding to both matrices an extra row whose entries are equal to the sum of the entries in the subset S gives a representation of a binomial B on a prolongation of T. □

In particular, this means that if the conjecture 4.1 of Sturmfels and Sullivant holds for the 3-Kimura model, then conjecture 4.6 also holds for this model for $n > 257$. Later we will significantly improve this estimation.

For the second modification of the conjecture 4.6 let us recall a few facts on toric varieties. Let T_1 and T_2 be two tori with lattices of characters given respectively by M_1 and M_2. Assume that both of them are contained in a third torus T with the character lattice M. The inclusions give natural isomorphisms $M_1 \simeq M/K_1$ and $M_2 \simeq M/K_2$, where K_1 and K_2 are torsion free lattices corresponding to characters that are trivial when restricted respectively to T_1 and T_2. The ideal of each torus (inside the big torus) is generated by binomials corresponding to such trivial characters. The points of T are given by semigroup morphisms $M \to C^*$. The points of T_i are those morphisms that associate 1 to each character from K_i. We see that the points of the intersection $T_1 \cap T_2$ are those morphisms $M \to \mathbb{C}^*$ that associate 1 to each character from the lattice $K_1 + K_2$. Of course the (possibly reducible) intersection Y is generated by the ideal corresponding to $K_1 + K_2$. This lattice may be not saturated, but Y contains a distinguished torus T', that is one of its connected components. If K' is the saturation of the lattice $K_1 + K_2$ then the characters of T' are given by the lattice M/K'. Let X_i be the toric variety that is the closure of T_i, and X' be the closure of T'. We call the toric variety X' the *toric intersection* of X_1 and X_2.

In the setting of 4.6 we conjecture the following:

Conjecture 4.8. *The toric variety $X(T)$ is the toric intersection of all the toric varieties $X(T_i)$.*

This conjecture differs from the previous one by the fact that we allow the intersection to be reducible, with one distinguished irreducible component equal to $X(T)$. We state this conjecture, because it can be checked using only the tori. As the biologically meaningful points are contained in the torus (see [CFS08]), this conjecture is of much importance for applications. Moreover, it is quite easy to check it for trees with small enough number of leaves using the computer programs. To explain it properly, let us consider the following general setting.

Assume that the tori T_i are associated to polytopes P_i and that T is just the torus of the projective space $\mathbb{P}^n \supseteq T_i$. Let A_i be a matrix whose columns represent vertices of the polytope P_i. The characters trivial on T_i or respectively binomials generating the ideal of T_i are exactly represented by integer vectors in the kernel of A_i. The characters trivial on the intersection are given by integer vectors in $\ker A_1 + \ker A_2$.

Note that the ideal of the toric intersection T' of the tori T_i in T is generated by binomials corresponding to characters trivial on T', that is by the saturation of $\ker A_1 + \ker A_2$. These binomials define a toric variety in $\mathbb{P}^n$. This variety is contained in the intersection (in fact it is a toric component) of the toric varieties that are the closures of T_i. The equality may not hold however, as the intersection might be reducible.

In conjecture 4.8 we have to compare two tori, one contained in the other. To do this, it is enough to compare their dimension, that is the rank of the character lattice. Let us note that the dimension of the intersection $T_1 \cap T_2$ is given by n minus the dimension (as a vector space) of $\ker A_1 + \ker A_2$, as it is equal to the rank of the lattice $\mathbb{Z}^n \cap (\ker A_1 + \ker A_2)$. To compute this dimension it is enough to compute the ranks of matrices A_1, A_2 and B, where B is a matrix obtained by putting A_1 under A_2 (that is, $\ker B = \ker A_1 \cap \ker A_2$). This can be done very easily using GAP ([GAP]).

The results obtained for small trees will be used in the following section.

4.3. **Main Results.** To support the conjecture 4.6 let us consider the case of binary model. This model is well understood [BW07], [CP07], [SS05]. Now we can prove the following:

Proposition 4.9. *Conjecture 4.6 holds for the binary model.*

Proof. We use the same notation as in the proof of proposition 4.7. From 4.2 we know that $\phi(\mathbb{Z}_2) = 2$. Let us consider any binomial of degree 2 for a claw tree with n leaves. This is given by a pair of matrices A_1, A_2 with 2 columns each. Let $A = A_1 - A_2$, where the difference uses the group law. We construct a subset S of the set of rows which gives a prolongation of the tree.

By permuting columns of A_2 we may assume that the entries in the first row of the matrix A are trivial. Let A' be the matrix obtained by deleting the first row of A. If we have a row 00 in A' then we are done, so assume there are only 01, 10 and 11. Notice that 01 and 10 cannot occur at all, as A_1 and A_2 would not have the same rows up to permutation. For $n > 3$ we can take twice 11 as a strict subset of the set of rows, summing up to zero in each column. □

From the proof above it follows that in fact to obtain the variety of the claw tree for the binary model it is enough to intersect three varieties corresponding just to three subdivisions. This subdivisions correspond to S containing exactly first and second row or first and third, or second and third row.

Now we prove the following conditional result for the 3-Kimura model:

Proposition 4.10. *If the conjecture 4.1 of Sturmfels and Sullivant holds then the conjecture 4.6 holds for $n > 8$.*

Proof. We use the same notation as in the previous proof, but instead of considering the matrix A (corresponding to a chosen binomial) with k columns and entries from $\mathbb{Z}_2 \times \mathbb{Z}_2$ we assume that it has $2k$ columns and entries from $\mathbb{Z}_2$. Let us note that the number of 1 in each row both in even and odd columns has to be even. This follows from the fact that rows of A are differences of rows that were equal up to permutation. This means that both projections from $\mathbb{Z}_2 \times \mathbb{Z}_2 \to \mathbb{Z}_2$ gave rows that were equal up to permutation. The difference of such vectors has got always an even number of 1.

Once again we may assume that the entries in the first row of A are trivial characters, that is they are equal to zero. Let A' be the matrix obtained by deleting the first row of A. For each subset of rows of A' we may consider a vector of length equal to the number of columns of A', whose entries are given by sums of characters from the subset. Note that this vector always has an even number of 1 both in even and odd columns. Because we assume conjecture 4.1, the matrix A' has got at most 8 columns. By Dirichlet's principle, if $n > 8$ then we can find two subsets of rows of A' that are not complements of each other, such that their sum vector is the same. If we take a symmetric difference of these subsets, we obtain a strict, nonempty set S of rows of A', summing up in each column to a trivial character. We add the first row of A to S or its complement, so that both sets have more than one element. Thus we obtain a subdivision of the set of rows

of A such that the given binomial is in the ideal of the tree corresponding to this division. □

For $n \leq 8$ we checked, using the computer programs Polymake, 4ti2, Macaulay2 and GAP, that the toric intersection of the tori of subdivisions gives the torus of the claw tree. We used the linear algebra described in the previous section. This proves that if the conjecture 4.1 holds for 3-Kimura model, then the conjecture 4.8 holds. Moreover, in all the checked cases it was enough to consider just two subdivisions.

To summarize, we know that for 3-Kimura model conjecture 4.6 implies both conjectures 4.8 and 4.1 and moreover conjecture 4.1 implies 4.8 and for $n > 9$ also conjecture 4.6.

Appendix

Here we present the precise results of the computations of Hilbert-Ehrhart polynomials for a few G-models, stated in 3.3. For each the first groups we considered the numbers of lattice points in consecutive dilations are given.

For the groups $\mathbb{Z}_8$, $\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$ and $\mathbb{Z}_9$ we computed only the Hilbert function and, as we could not check the normality, we do not know if it is equal to Hilbert-Ehrhart polynomial.

Models for $G = H = \mathbb{Z}_3$**.**

dilation	*snowflake*	*3-caterpillar*
1	243	243
2	21627	21627
3	903187	904069
4	21451311	21496023
5	330935625	331976637
6	3647265274	3662146270
7	30770591364	30920349834
8	209116329075	210269891871
9	1189466778457	1196661601837
10	5831112858273	5868930577941
11	25205348411361	25377886917819

Models for $G = H = \mathbb{Z}_2 \times \mathbb{Z}_2$ **(3-Kimura).**

dilation	*snowflake*	*3-caterpillar*
1	1024	1024
2	396928	396928
3	69248000	69324800
4	5977866515	5990170739
5	291069470720	291864710144
6	8967198289920	8995715702784

Models for $G = H = \mathbb{Z}_4$**.**

dilation	*snowflake*	*3-caterpillar*
1	1024	1024
2	396928	396928
3	69248000	69324800
4	6122557220	6138552524
5	310273545216	311525688320
6	10009786400352	10062179606880

Models for $G = H = \mathbb{Z}_5$**.**

dilation	*snowflake*	*3-caterpillar*
1	3125	3125
2	3834375	3834375
3	2229584375	2230596875
4	640338121875	642089603125

Models for $G = H = \mathbb{Z}_7$**.** In this case the first three dilations of the polytopes have the same number of points. The numbers of points in fourth dilations were too big to obtain precise results. Hence we computed only the numbers of points mod 64, which is sufficient to prove that the Hilbert-Ehrhart polynomials are different.

dilation	*snowflake*	*3-caterpillar*
1	16807	16807
2	117195211	117195211
3	423913952448	423913952448
4	$\equiv 54 \mod 64$	$\equiv 14 \mod 64$

Models for $G = H = \mathbb{Z}_8$.

dilation	*snowflake*	*3-caterpillar*
1	32768	32768
2	454397952	454397952
3	3375180251136	3375013036032

Models for $G = H = \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$.

dilation	*snowflake*	*3-caterpillar*
1	32768	32768
2	454397952	454397952
3	3375180251136	3375013036032

Models for $G = H = \mathbb{Z}_9$.

dilation	*snowflake*	*3-caterpillar*
1	59049	59049
2	1499667453	1499667453
3	20938605820263	20937202945056

References

[AR08] Elizabeth S. Allman and John A. Rhodes, *Phylogenetic ideals and varieties for the general Markov model*, Advances in Applied Mathematics **40(2)** (2008), 127–148.

[BDW09] Weronika Buczyńska, Maria Donten, and Jarosław A. Wiśniewski, *Isotropic models of evolution with symmetries*, Contemporary Mathematics **496** (2009), 111–132.

[BW07] Weronika Buczyńska and Jarosław A. Wiśniewski, *On geometry of binary symmetric models of phylogenetic trees*, J. Eur. Math. Soc. **9(3)** (2007), 609–635.

[CFS08] M. Casanellas and J. Fernandez-Sanchez, *Geometry of the Kimura 3-parameter model*, Advances in Applied Mathematics **41(3)** (2008), 265–292.

[CP07] J. Chifman and S. Petrović, *Toric ideals of phylogenetic invariants for the general group-based model on claw trees $k_{1,n}$*, Proceedings of the 2nd international conference on Algebraic biology (2007), 307–321.

[DK09] Jan Draisma and Jochen Kuttler, *On the ideals of equivariant tree models*, Mathematische Annalen **344(3)** (2009), 619–644.

[ERSS04] N. Eriksson, K. Ranestad, B. Sturmfels, and S. Sullivant, *Phylogenetic algebraic geometry*, Projective Varieties with Unexpected Properties; Siena, Italy (2004), 237–256.

[GAP] *GAP — Groups, Algorithms, and Programming, Version 4.4.10*, The GAP Group, (http://www.gap-system.org), 2007.

[GJ00] Ewgenij Gawrilow and Michael Joswig, *Polymake: a Framework for Analyzing Convex Polytopes*, Polytopes — Combinatorics and Computation (Gil Kalai and Günter M. Ziegler, eds.), Birkhäuser, 2000, pp. 43–74.

[Ilt10] Nathan Ilten, *Deformations of Rational Varieties with Codimension-One Torus Action*, Doctoral Thesis, FU Berlin, 2010.

[Kub10] Kaie Kubjas, *Hilbert polynomial of the Kimura 3-parameter model*, arXiv:1007.3164v1 [math.AC] (2010).

[Mic10] Mateusz Michałek, *Algebraic varieties representing group-based Markov processes on trees*, arXiv:1004.3012v1 [math.AG] (2010).

[PS05] Lior Pachter and Bernd Sturmfels, *Algebraic Statistics for Computational Biology*, Cambridge University Press, 2005.

[SS05] Bernd Sturmfels and Seth Sullivant, *Toric ideals of phylogenetic invariants*, J. Comput. Biology **12** (2005), 204–228.

[Sul07] Seth Sullivant, *Toric fiber products*, J. Algebra **316** (2007), 560–577.

[tt] 4ti2 team, *4ti2—a software package for algebraic, geometric and combinatorial problems on linear spaces*, www.4ti2.de.

Maria Donten-Bury
`marysia@mimuw.edu.pl`
Mathematics Institute, University of Warsaw, Banacha 2, 02-097 Warszawa, Poland

Mateusz Michałek
`wajcha2@poczta.onet.pl`
Mathematical Institute of the Polish Academy of Sciences, Św. Tomasza 30, 31-027 Kraków, Poland
Institut Fourier, Universite Joseph Fourier, 100 rue des Maths, BP 74, 38402 St Martin d'Hères, France