1. Research Plan: Nature of Research and Significance

1.1. **Introduction.** A *phylogenetic tree* (or *phylogeny*) is a biological family tree. Mathematically it is a *graph* — a collection of items connected by branching edges — which summarizes the relations of evolutionary descent between different species, organisms, or genes. Phylogenetic trees are useful tools for reasoning about events which may have occurred in the evolutionary history of an organism. The immediate impact of such tools is in the field of natural history. The bigger human impact, however, is to biogenetics. This knowledge allows us to postulate that similar adaptations are probably inherited features of a common ancestor, rather than convergent evolution due to a common advantage. If species that share a biological trait can all be shown to have a common origin, then we suspect that trait has a common genetic basis and look for its genes among the shared portions of their genomes. Discovering how to better link genes and their expression in humans will let us read the DNA code and search for disease-causing errors.

Current efforts to reconstruct phylogenies demand inferences from thousands of DNA sequences [11] [3]. Access to bigger data sets can sometimes lead to even more questions. Numerous studies have claimed that Amborella (found in New Caledonia) is the solitary sister to the rest of flowering plants. Yet, several studies have shown that the first branch is composed of Amborella and Nymphaeales (water lilies). Figure 1 shows three of the current theories of angiosperm evolution.

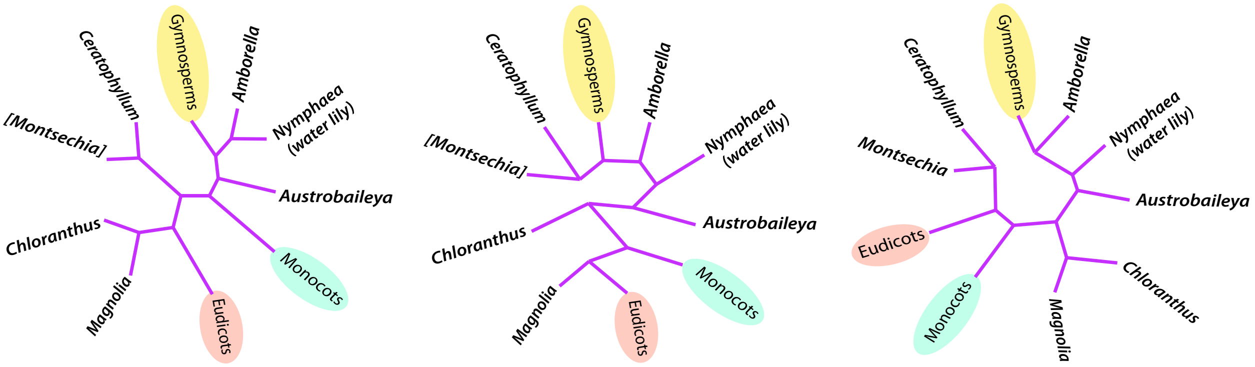

Figure 1: **Phylogenetic trees for angiosperms (flowering plants), for each of which the gymnosperms are the root ancestor. Left-to-right these are based on results in [1], [12] and [7]. The latter is the only source where the fossil Montsechia appears, but we add it to the first two as a sister to extant Ceratophyllum for comparison.**

Historically many theoretical methods of reconstructing a phylogenetic tree have been used, with varying speed and accuracy. Recently a class of algorithms known as *distance-based methods* have been developed. Given an input list of dissimilarities (genetic distances) between the pairs of species in our starting set, we want to place those species as the *leaves* on a phylogenetic tree that represents their most likely evolutionary history. In this proposal, we initially focus on the distance-based method known as the *balanced minimal evolution* (BME) method, which creates a tree whose total length (sum of branch lengths) is minimal.

It turns out that solving this problem is equivalent to finding the optimal corner, or *vertex*, of a large, high-dimensional shape called a *polytope*. The number of possible phylogenetic trees is the number of vertices of this *BME polytope*, and this number is given by a double factorial formula. We show the first example in Figure 2, but the number of vertices grows rapidly: there are 654,729,075 possible trees (and thus that same number of vertices) when the number of species is 12. Thus the direct route of simply testing each possible tree is not feasible. Instead there is a branch of mathematics called *linear programming* which allows us to take shortcuts to the answer, but only if we know the *linear inequalities* that define the sides, or *facets*, of our polytope. Before our research the facet linear inequalities (such as $2x + 3y \leq 7$) that define the BME polytope were completely unknown.

1.2. Terminology.

- A *clade* is a sub-tree of a phylogenetic tree which has a common ancestor (branch point) for all its leaves.
- A *cherry* is a clade with only two leaves.
- A pair of *intersecting cherries* $\{a, b\}$ and $\{b, c\}$ have intersection in one leaf b, and thus cannot exist both on the same tree.
- A *caterpillar* is a tree with only two cherries.
- A *split* of the set of n leaves for our phylogenetic trees is a partition of the leaves into two parts, one part called S with m leaves and another with the remaining $n - m$ leaves. A tree *displays* a split if each part makes up the leaves of a *clade*. See Figure 2.

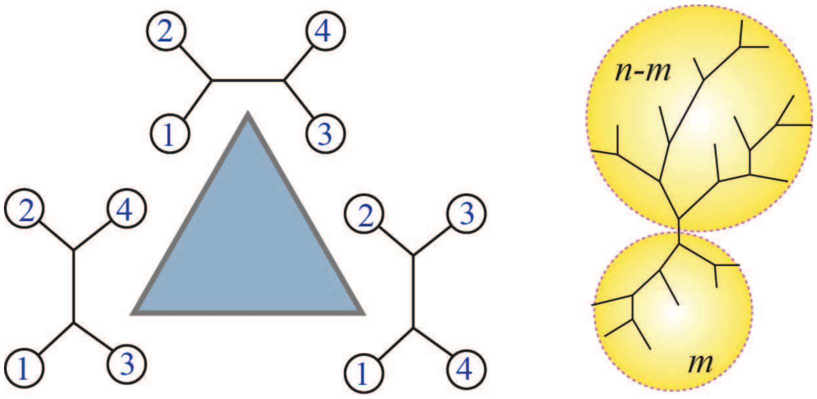

Figure 2: **The BME polytope $\mathcal{P}_4$ is a triangle, with vertices labeled by three caterpillar trees. On the right is a tree displaying a split into two circled clades.**

1.3. **Collaboration.** The research proposed here is one part of a larger project that includes collaborators from the University of Akron, the University of Western Michigan, Clemson University and the University of Kentucky. (See list of collaborators.)

2. Goals and Objectives

There are two interrelated mathematical objectives: (1) to understand more of the facet structure of the BME polytope; and (2) to use that structure in fast, accurate algorithms for reconstructing phylogenetic trees. Early experiments carried out by the PI and students are quite promising for both of these objectives! Our initial biological objective, with collaborators, is to answer the question of flowering plant evolution.

2.1. **Objective 1: Describing the Geometric Combinatorics of Evolution.** Recently we have uncovered features of the BME polytope that should allow both greater mathematical understanding and practical leverage.

2.1.1. *What we found: preliminary results for the BME polytope.* We have been able to describe many new faces, especially facets, of the n^{th} balanced minimal evolution polytope $\mathcal{P}_n$. (Our results are listed in columns 5–7 of Table 1.)

The PI and students have shown in [6] that any pair of intersecting cherries corresponds to a facet of $\mathcal{P}_n$. Each pair of cherries with leaves $\{a, b\}$ and $\{b, c\}$, there is a facet of $\mathcal{P}_n$ whose vertices correspond to trees that have one of those two cherries. See Figure 3.

In addition, any caterpillar tree with fixed ends corresponds to a facet of $\mathcal{P}_n$. Thus for each pair of species there is a facet of $\mathcal{P}_n$ whose vertices correspond to trees which are caterpillars with this pair as far apart as possible. Also shown in [6]: for $n = 5$, for each necklace of five leaves there is a corresponding facet which is combinatorially equivalent to a simplex.

New work: Splits and facets. More recently we have postulated the existence of an exponentially increasing family of *split-facets.* We discovered that the trees which display a given split form a face. The inequality for a split-face is shown in Table 1. We have proven that it becomes an equality for the trees displaying the split; and becomes a strict inequality for all other trees. Our objective is to first completely prove that the large split-faces are all facets by showing that they are truly sides of the polytope. Then we will generalize the concept to find more facets and faces.

The Splitohedron. The split-faces, intersecting-cherry faces, and caterpillar facets together outline a new shape that lies within a high-dimensional hypercube. We decrease the dimension of that shape using basic equalities obeyed by all the phylogenetic trees, and this defines a new polytope: the *splitohedron*.

For a given phylogenetic tree it is straightforward to count how many distinct facets of the splitohedron that tree belongs to. If that number is as large as the dimension, we know that the tree lies at a true vertex of the polytope. We have shown that (for $n \leq 12$) the splitohedron contains among its vertices all the possible phylogenetic trees. Therefore if a BME linear program is optimized in the splitohedron at a valid tree vertex, it is also optimized in the BME polytope. Thus by finding further facets we will improve this theorem, hopefully to a version that holds for all $n > 12$.

number of species	dim. of $\mathcal{P}_n$	vertices of $\mathcal{P}_n$	facets of $\mathcal{P}_n$	facet inequalities (classification)	number of facets	number of vertices in facet
3	0	1	0	-	-	-
4	2	3	3	$x_{ab} \geq 1$	3	2
				$x_{ab} + x_{bc} - x_{ac} \leq 2$	3	2
5	5	15	52	$x_{ab} \geq 1$ (caterpillar)	10	6
				$x_{ab} + x_{bc} - x_{ac} \leq 4$ (intersecting-cherry)	30	6
				$x_{ab} + x_{bc} + x_{cd} + x_{df} + x_{fa} \leq 13$ (cyclic ordering)	12	5
6	9	105	90262	$x_{ab} \geq 1$ (caterpillar)	15	24
				$x_{ab} + x_{bc} - x_{ac} \leq 8$ (intersecting-cherry)	60	30
				$x_{ab} + x_{bc} + x_{ac} \leq 16$ $(3,3)$-split	10	9
n	$\binom{n}{2} - n$	$(2n-5)!!$	?	$x_{ab} \geq 1$ (caterpillar)	$\binom{n}{2}$	$(n-2)!$
				$x_{ab} + x_{bc} - x_{ac} \leq 2^{n-3}$ (intersecting-cherry)	$\binom{n}{2}(n-2)$	$2(2n-7)!!$
				$x_{ab} + x_{bc} + x_{ac} \leq 2^{n-2}$ $(m,3)$-split, $m > 3$	$\binom{n}{3}$	$3(2n-9)!!$
				$\sum_S x_{ij} \leq (m-1)2^{n-3}$ $(m, n-m)$-split S, $m > 2, n > 5$	$2^{n-1} - \binom{n}{2}$ $-n-1$	$(2(n-m)-3)!!$ $\times(2m-3)!!$

Table 1: Technical statistics for the BME polytopes $\mathcal{P}_n$. The first four columns are found in [9] and [8]. Our new results are in the last 3 columns. The inequalities are given for any $a, b, c, \dots \in [n]$. Note that for $n = 4$ the three facets are described twice: our inequalities are redundant. The last two types of facet listed are shown to be faces, and believed to be top-dimensional.

2.2. **Objective 2: Algorithms for Tree Reconstruction.** The number of facets of the BME polytope seems to grow exponentially and so a simple implementation of linear programming is unlikely. What we have instead is the Splitohedron just described: a *relaxation* of the BME polytope. A relaxation of a polytope is a simpler, enveloping polytope. Our facet inequalities have many desirable features. For instance, all 654,729,075 phylogenetic trees on 12 species are recovered in the 54 dimensional polytope defined by our list of facet inequalities.

However our splitohedra also contain many invalid vertices. Thus our algorithm for finding the BME trees will work as follows: (1) Given the pairwise genetic distances for a set of n species, we automatically generate the facet inequalities for the n^{th} splitohedron. (2) We use these inequalities as input for standard linear programming, which may return a valid or an invalid answer. Note that the answers here are trees encoded as vectors.

(3) If we get a valid result, the algorithm is finished. Otherwise, for each of the invalid components we introduce additional inequalities to restrict the component away from its original value. (This is the *branching stage* of a *branch-and-bound* algorithm.) Each new pair of inequalities yields a couple of smaller polytopes, and we repeat the process for both. (4) The algorithm terminates when the (rerun) linear program solver returns valid solutions, and we choose the minimal answer from that smaller set.

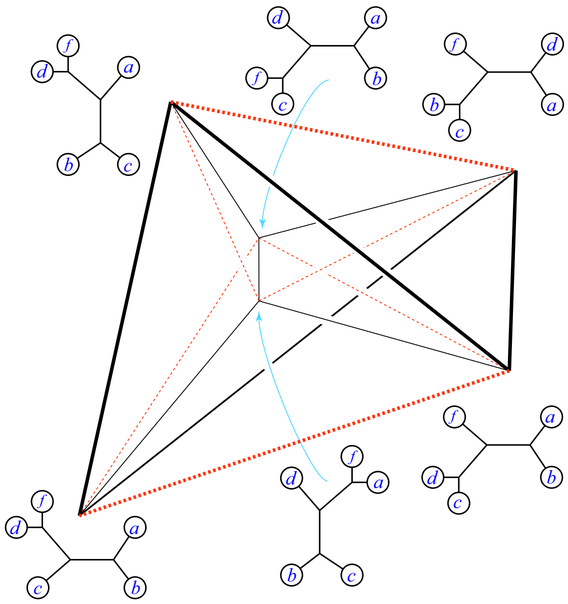

Figure 3: **4-dimensional facet of $\mathcal{P}_5$ with each vertex labeled by a tree which contains one of two intersecting cherries:** $\{a, b\}$ **and** $\{b, c\}$**.**

2.3. **Additional Objectives: Applications and Extensions.** With our collaborators, we plan to apply our new methods to open questions about the evolutionary relationships among flowering plants. Collaborator Dr. Todd Barkman at UWM describes the planned work as follows:

“An abundance of data exist that our methods are particularly applicable to. Specifically, we will make use of complete mitochondrial genome sequence datasets, complete chloroplast genome data sets, as well as nuclear transcriptome/genome datasets for hundreds of taxa. We view these as the perfect empirical data sets by which to showcase the methodological advances our work should achieve, partly due to the computational efficiency that [distance-based methods] provide and is needed for large datasets such as these.”

We also envision extending our methods to phylogenetic networks which include hybridization and lateral gene transfer. The graphs in these cases will no longer be trees. In [10] Levy and Pachter suggest utilizing the *graph-associahedra* as defined by Devadoss and Carr in [2]. These are polytopes whose facets correspond to certain subgraphs. Their vertices and the vertices of the related *multiplihedra* use powers of 3 as shown in [4]. We would like to generalize the BME polytope along these lines, and pursue connections to polytopes such as the *permutahedron* and the *permutoassociahedron*.

2.4. **Procedures.** The two main procedures for our planned investigations are computational experiments and theoretical proof of the patterns we find.

Computational experiments. We are using the programming platforms Polymake (free) and Matlab (licensed by UA), as well as potentially other linear programming solvers. We create input data from sets of trees, and then Polymake outputs the inequalities and sets of vertices that correspond to each facet. Then comes the difficult work of creatively finding a pattern in the output that applies to any size input set. Students have been able to participate in all these steps.

Theoretical proofs. The proofs are constructive: first we show that the trees in the facet satisfy *sharply* (as an equality) the inequality strictly obeyed by all other trees. Then we inductively describe a *flag* (sequence of included faces) that shows a codimension of 1 for our facet.

More computation. We are coding (with Matlab) a program that will generate all the facet inequalities for n species. Then we plan to use that program to test our branch-and-bound algorithm on hand-generated examples. Again the goal is to find patterns: this time we will look for ways to guarantee accuracy and improve speed.

Future application. In the larger research project, the next step is to work with Todd Barkman at the University of Western Michigan to run our program on data from flowering plants.

3. Time Table

3.1. **Spring 2016.** We will continue the investigations already begun on the first main objective: finding as many facets, faces and edges of the BME polytope as possible. As a consequence we will be able to generate a more complete picture of how they intersect with and contain each other. This project is especially suited to our undergraduate and master's degree candidates.

3.2. **Summer 2016.** We will refine the new relaxation of the BME linear programming, adding precision to our Splitohedron method. Also we will experiment on simulated data, testing our branch-and-bound algorithm for speed and accuracy and using the results to improve it.

3.3. **Fall 2016.** We will submit our updated proposal to the National Science Foundation and the National Institutes of Health: the Department of Mathematical Sciences and the National Institute of General Medical Sciences (DMS/NIGMS) Joint Initiative to Support Research at the Interface of Biological and Mathematical Sciences. We will extend results beyond the phylogenetic trees to networks showing horizontal gene transfer, hybridization, and genetic drift.

4. Expected Results

We would eventually like a complete description of $\mathcal{P}_n$, both combinatorially and geometrically. On the combinatorial side we would like to know which sets of vertices are those of a facet, and what other polytopes those facets are equivalent to. On the geometrical side we would like to know how to quickly find the list of facet inequalities that describe $\mathcal{P}_n$.

We also postulate that: (1) partitions which refine the split correspond to sub-faces of the split-facet; and (2) giving two or more splits which the trees must all display simultaneously corresponds to a sub-face of a split-facet.

We hypothesize that, for an input list of pair-wise genetic distances, linear programming over the spitohedron is a fast, consistent and stochastically accurate method for finding the BME polytope. This is based on testing with small leaf sets (and in long-branch situations) as well as on the features of the splitohedron that we have uncovered.

Intersections of facets: Every facet defined by a collection of caterpillar trees intersects all the other facets from caterpillar trees. That is, for any two of these facets (one for $\{a, b\}$ and another for $\{c, d\}$) there is face which the two facets have in common. It is the face with vertices given by caterpillar trees that have either of these two pairs of cherries: ($\{a, c\}$ and $\{b, d\}$), or ($\{a, d\}$ and $\{b, c\}$). See Figure 4 where the intersection is an edge.

Eickmeyer et.al. show in [5] that the caterpillar trees often have the worst chances for being correctly reconstructed by existing algorithms. It is hypothesized (about another distance-based method) that for $n > 6$, the caterpillar tree is "the most difficult BME tree topology for the

NJ algorithm to reproduce." Since our splitohedron obeys caterpillar-tree-based inequalities, we believe it to have a better chance of finding these solutions.

5. Publication and Presentation

My students and I have already been actively disseminating the initial stages of this research: we have published a paper [6] and presented at the meeting of the American Mathematical Society (AMS), at Loyola University Chicago this past October. I have also presented this research via four talks in the UA math department algebra and combinatorics seminar.

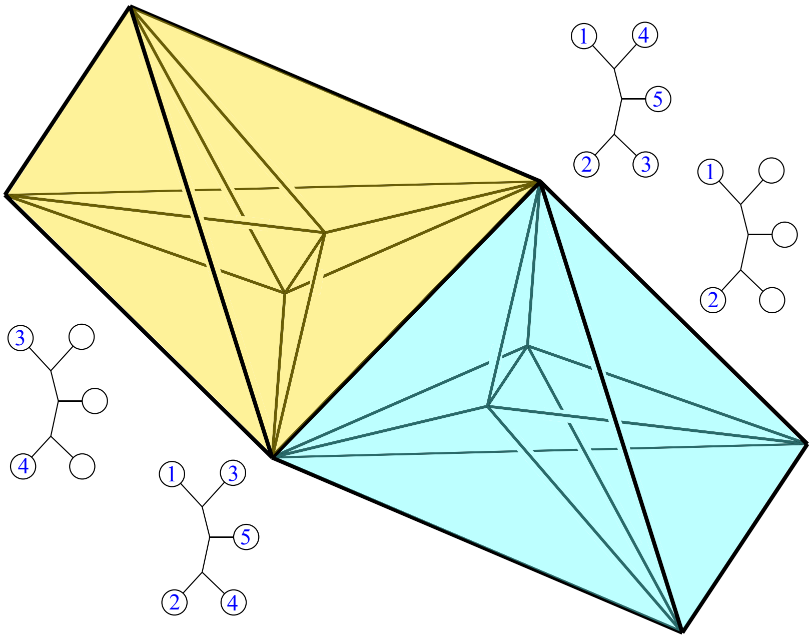

Figure 4: **Two intersecting caterpillar facets of BME(5).**

Several more presentations are planned: a refereed proceedings and competitive conference presentation at Formal Power Series and Combinatorics (FPSAC 2016) in Vancouver (the premier annual combinatorics meeting); an invited talk at a proposed Mini-symposium at the Society for Industrial and Applied Mathematics (SIAM) 2016 Annual Meeting/ Conference on the Life Sciences in Boston; and invited participation at a July 2016 workshop at the National Institute for Mathematical and Biological Synthesis (NIMBioS) in Knoxville.

We also plan to publish sequel(s) to our paper mentioned above, perhaps also in the Journal of Mathematical Biology. All these methods of dissemination are crucially important to achieving the requirements for RTP and merit in the mathematics department. One further method of dissemination we have also begun is a website entitled "Encyclopedia of Combinatorial Polytope Sequences." The database is hosted on the U.A. website at `http://www.math.uakron.edu/~sf34/hedra.htm`. There are many completed entries already, from the classic sequences to new discoveries.

References

[1] Todd J. Barkman, Gordon Chenery, Joel R. McNeal, James Lyons-Weiler, Wayne J. Ellisens, Gerry Moore, Andrea D. Wolfe, and Claude W. dePamphilis. Independent and combined analyses of sequences from all three genomic compartments converge on the root of flowering plant phylogeny. *Proceedings of the National Academy of Sciences*, 97(24):13166–13171, 2000.

[2] Michael P. Carr and Satyan L. Devadoss. Coxeter complexes and graph-associahedra. *Topology Appl.*, 153(12):2155–2168, 2006.

[3] Francesca D. Ciccarelli, Tobias Doerks, Christian von Mering, Christopher J. Creevey, Berend Snel, and Peer Bork. Toward automatic reconstruction of a highly resolved tree of life. *Science*, 311(5765):1283 – 1287, March 2006.

[4] Satyan Devadoss and Stefan Forcey. Marked tubes and the graph multiplihedron. *Algebr. Geom. Topol.*, 8(4):2081–2108, 2008.

[5] K. Eickmeyer, P. Huggins, L. Pachter, and R. Yoshida. On the optimality of the neighbor-joining algorithm. *Algorithms for Molecular Biology*, 3(5), 2008.

[6] S. Forcey, L. Keefe, and W. Sands. Facets of the balanced minimal evolution polytope. *Journal of Mathematical Biology*, 2015.

[7] Bernard Gomez, Vronique Daviero-Gomez, Clment Coiffard, Carles Martn-Closas, and David L. Dilcher. Montsechia, an ancient aquatic angiosperm. *Proceedings of the National Academy of Sciences*, 112(35):10985–10988, 2015.

[8] David C. Haws, Terrell L. Hodge, and Ruriko Yoshida. Optimality of the neighbor joining algorithm and faces of the balanced minimum evolution polytope. *Bull. Math. Biol.*, 73(11):2627–2648, 2011.

[9] P. Huggins. Polytopes in computational biology. *Ph.D. Dissertation, U.C. Berkeley*, 2008.

[10] D. Levy and Lior Pachter. The neighbor-net algorithm. *Advances in Applied Mathematics*, 47:240–258, 2011.

[11] D.R. Maddison, K.S. Schulz, and W.P. Maddison. The tree of life web project. *Linnaeus Tercentenary: Progress in Invertebrate Taxonomy. Zootaxa 1668:1-766*, pages 19 – 40, 2010.

[12] Cynthia M. Morton. Newly sequenced nuclear gene (xdh) for inferring angiosperm phylogeny. *Ann. of Missouri Botanical Garden*, 98(1):63–89, 2011.

6. Feasability

The research described herein (both finished and proposed) is an application of the PI's experience in geometric combinatorics to the relatively new research program on phylogenetics that is being developed. Besides the recent paper [6], the PI has multiple publications in the area of geometric combinatorics. These include a series of papers proving the polytope structures for an extended family of graph-related combinatorial sets. Titles include (see CV for publication information):

- Convex Polytopes from Nested Posets.
- Pseudograph associahedra.
- Geometric combinatorial algebras: cyclohedron and simplex.
- Marked tubes and the graph multiplihedron.
- Convex Hull Realizations of the Multiplihedra.
- Quotients of the multiplihedron as categorified associahedra.

Most recently the PI was funded for two consecutive years as a Young Investigator of the Mathematical Sciences Program run by the National Security Agency. The second year of this funding actually supported the initial stages of the research described in this proposal.

6.1. **Collaborators.** Dr. Matt Macauley (mathematics, Clemson) and students, and Logan Keefe (UA) are collaborating on objective 1. William Sands (UA) is collaborating on objective 2. Dr. Todd Barkman (biology, UWM) is our contact for biological applications for flowering plants. Dr. Ruriko Yoshida (statistics, UKY) and Dr. Terrell Hodge (mathematics, UWM) are also collaborators on the larger NSF/NIH proposal.

6.2. **Funding status.** There is currently no internal (start-up) funding available for the PI. There is also currently no external funding available for Summer 2016. We have submitted a grant proposal to the NSF/NIH Joint Initiative to Support Research at the Interface of Biological and Mathematical Sciences, which is pending approval.

7. Previous, current and future efforts towards external funding.

7.1. **Previous.** In the summer of 2014 and 2015, the PI and students were partially supported by a Young Investigator grant from the Mathematical Sciences Program of the National Security Agency. After the first year of that grant, during which we pursued the theoretical foundations of geometric combinatorics, the PI changed directions to focus more on the polytopes and applications described in this current research proposal, and the grant was renewed for summer 2015.

7.2. **Current.** Now that the above-mentioned support is ending, we have joined with a group of collaborators from multiple universities (Western Michigan, Clemson, and the University of Kentucky) to attack several related phylogenetic questions. We turned in (on Sept. 15, 2015) an initial proposal for this larger project, including an early version of the research proposed here, to a joint program of the National Science Foundation and the National Institutes of Health: the Department of Mathematical Sciences and the National Institute of General Medical Sciences (DMS/NIGMS) Joint Initiative to Support Research at the Interface of Biological and Mathematical Sciences. If that proposal is successful it would cover 1.5 summer months of salary for the PI starting in 2016, and support graduate students as well.

7.3. **Future.** In the months since we submitted our proposal to the DMS/NIGMS program, we have already made progress—reported on here in this proposal. Thus, if due to restricted federal funds we are not successful in the 2015-2016 cycle, we plan to improve our proposal with new results and resubmit to the NSF and/or NIH in Fall 2016.