The University of Akron
IdeaExchange@UAkron

Honors Research Projects

The Dr. Gary B. and Pamela S. Williams Honors College

Spring 2016

New Facets of the Balanced Minimal Evolution Polytope

Logan Keefe
ltk7@zips.uakron.edu

NEW FACETS OF THE BALANCED MINIMAL EVOLUTION POLYTOPE

A Thesis

Presented to

The Graduate Faculty of The University of Akron

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

Logan Keefe

May, 2016

NEW FACETS OF THE BALANCED MINIMAL EVOLUTION POLYTOPE

Logan Keefe

Thesis

Approved: Accepted:

_____   _____
Advisor   Dean of the College
Dr. Stefan Forcey   Dr. John Green

_____   _____
Faculty Reader   Dean of the Graduate School
Dr. J.P. Cossey   Dr. Chand Midha

_____   _____
Faculty Reader   Date
Dr. Hung Nguyen

_____
Department Chair
Dr. Timothy Norfolk

ABSTRACT

The balanced minimal evolution (BME) polytope arises from the study of phyloge-

netic trees in biology. It is a geometric structure which has a variant for each natural

number n. The main application of this polytope is that we can use linear program-

ming with it in order to determine the most likely phylogenetic tree for a given genetic

data set. In this paper, we explore the geometric and combinatorial structure of the

BME polytope. Background information will be covered, highlighting some points

from previous research, and a new result on the structure of the BME polytope will

be given.

ACKNOWLEDGEMENTS

I would like to thank Dr. Forcey for all of the help and education he has given me over my time at The University of Akron, especially for choosing me to work on research with him and for being my thesis adviser. I would also like to thank Dr. Cossey and Dr. Nguyen for taking time to read my thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Figure 1.1: A naive and simplified version of a phylogenetic tree. Note the branching structure between the different taxa, as well as the interior nodes, highlighted with black dots. The terminal points, represented by white circles, are often called leaves of the tree and they represent the taxa in our case.

Trees are a natural choice for representing the relationships between taxa, with their use dating back to Darwin. Our goal with respect to trees is not one of the trees themselves, but of data that can be used to build them. A question is: given a set of distances between taxa, how can we construct the most accurate tree to fit this data?

The method we consider is the minimum-evolution method, so called due to the assumption that the tree with the smallest sum of branch lengths is the tree that

is the most likely to be correct [4]. From Pauplin's work [3], it follows that finding the tree which will minimize the distances is equivalent to minimizing the dot product of the vector of genetic distances with the corresponding vector of distances on the tree over the set of tree topologies, following a special metric. Since the genetic distances are constant and the topology of the tree can vary, this dot product gives a linear function.

In linear optimization, the function tends to be easy to describe while the domain can be quite the opposite. So, since the problem is reducible to minimizing a linear function, it is quite useful to study the space that we are minimizing over. We call this space the balanced minimal evolution (BME) polytope.
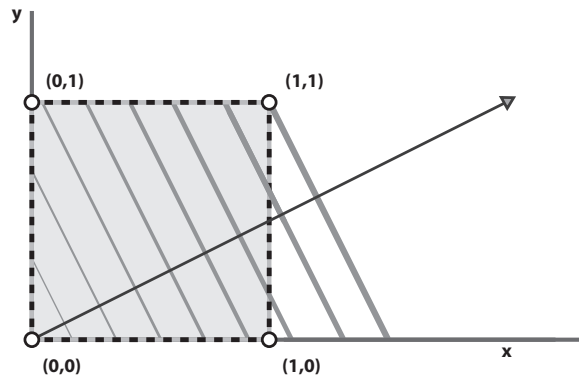
Figure 1.2: A visual representation of optimizing a linear function over a polytope.

CHAPTER II

POLYTOPES

2.1   Introduction to Polytopes
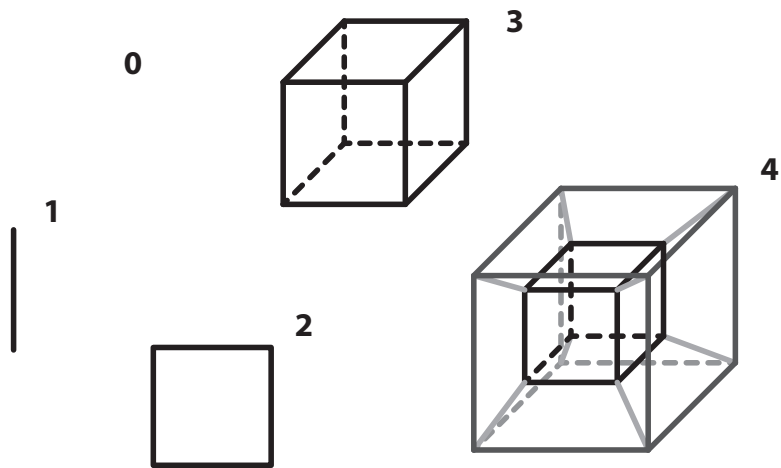
Figure 2.1: A set of images of polytopes in varying dimensions.

To understand what the BME polytope is, we must first understand what a polytope is. Some examples of polytopes include polygons, cubes, tetrahedrons, pyramids, and some higher dimensional figures such as a tesseract. There are several available definitions of what a polytope is, both combinatorial and geometrical. However, one in particular is used directly in the definition of the BME polytope, so that will be our introduction. Colloquially speaking, one way to define a polytope is a finite set of points which have been shrink wrapped. One might remember putting

rubber bands around pegs on a board. A polytope is a higher dimensional analogue of these types of figures. More formally speaking, a polytope is a convex hull of finitely many points in a Euclidean space. The definition of convex hull is as follows:

A set Y is said to be *convex* if for any points $a, b \in Y$, every point on the straight line segment joining them is also in Y. The *convex hull* of a set of points X in Euclidean pace is the smallest convex set containing X.
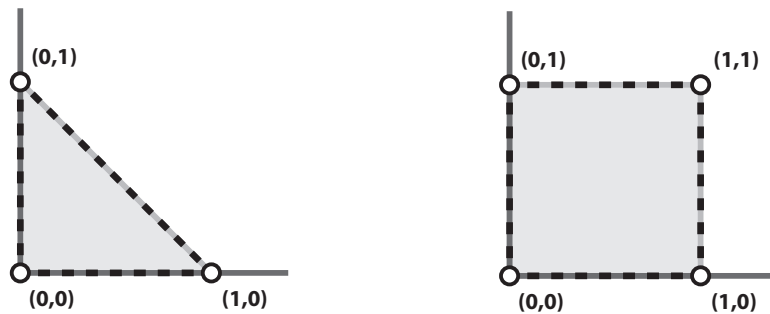
Figure 2.2: Two different convex hulls of slightly different sets of points. Notice how the shape changes with the addition of a point.

An easy example would be the set $\{(0,0), (1,0), (0,1), (1,1)\}$. As one can see, these points form a square. And the convex hull is exactly that. Note that we could, for example, add the point $(\frac{1}{2}, \frac{1}{2})$ to the set and we would get the exact same figure since this point is contained within the convex hull.

Another geometric definition of a polytope utilizes half-spaces, or linear inequalities. If we take finitely many linear equalities such that the set of points which obey all of them is bounded, we have a polytope. Any polytope given by a convex hull can also be given in this manner, as well as the other way around.

4

If we cannot remove a point v from the set we take the convex hull of without changing the convex hull itself, we will call v a *vertex* of the polytope. For example, each corner of a triangle is a vertex.

If there exists a linear inequality such that every point in the polytope satisfies it, we call the set of points that satisfy the equality portion a *face* of the polytope. Each face of a polytope is itself a polytope. For example, all of the corners and edges of an octagon are faces of the octagon.

The *dimension* of a polytope is the dimension of the smallest Euclidean space which could contain it. For example, the dimension of a pentagon is 2.

A *facet* of a polytope is a face of the polytope with dimension one less than that of the polytope. For example, a square is a facet of a cube.

Something that is important to note is that a polytope can also be described combinatorially as a partially ordered set, or more specifically a lattice. Each polytope is made up of smaller polytopes. Take for example a cube. It has square sides, which have lines making their sides, which have points making their sides. Each one of these is a progressively lower dimensional polytope. We will discuss this in more detail soon.

CHAPTER III

TREES AND COMBINATORICS

3.1   Some Combinatorial Ideas

A *partially ordered set*, or poset, is a set A together with a binary relation called a

partial order. A familiar partial order is the less than or equal to relation on the

real numbers, so that symbol is often used for partial orders in general. A partial

order must be reflexive, antisymmetric and transitive. A relation $\leq$ on A is *reflexive*

if $a \leq a$ for all $a \in A$. It is *antisymmetric* for $a, b \in A$ if when $a \leq b$ and $b \leq a$ we

have $a = b$. Finally, it is *transitive* for $a, b, c \in A$ if when $a \leq b$ and $b \leq c$ we have

$a \leq c$. An example of a poset is the real numbers ordered by less than or equal to

($\leq$), which is why the use of that symbol is common for posets in general.

A *lattice* is a special type of poset where any two elements have a unique least

upper bound and greatest lower bound. A common example is the natural numbers

ordered by divisibility. Here, the least upper bound of two numbers would be the

least common multiple, and the greatest lower bound would be the greatest common

divisor.

A *graph* is a set V of vertices together with a set E of edges such that every

element in E is of the form $\{x, y\}$ such that $x, y \in V$. These objects are distinct from

graphs of functions, and have applications in describing how objects are connected to one another.

A graph can be interpreted visually as some dots with lines drawn between dots. In practical problems, graphs almost always represent a series of connections. The nodes could represent cities and the edges roads, or the nodes could be computers and the edges network connections. In our case, the nodes represent taxa or branching points.
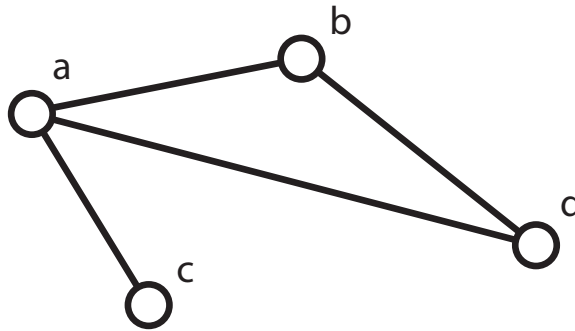
Figure 3.1: An image of the graph with vertex set $\{a, b, c, d\}$ and edge set $\{\{a, b\}, \{a, c\}, \{a, d\}, \{b, d\}\}$. We generally represent graphs pictorially since it makes them far easier to work with. Here, $\{\{a, b\}, \{b, d\}, \{d, a\}\}$ forms a cycle, so the graph is not a tree.

A *cycle* is a sequence in the edge set of the graph of the form

$$\{(x_1, x_2), (x_2, x_3), ..., (x_N, x_1)\}$$

A *tree* is a graph with no cycles. Visually, the graph has no loops in it and could be straightened out to look like a tree.

A *tree topology* is the underlying connectivity structure of a tree without regarding labelling.

We say a tree is *binary* if each vertex is either a member of one edge or three edges. We will call vertices associated with one edge *leaves* and vertices associates with three edges *nodes*. The graphs we work with will all be binary trees due to the nature of evolutionary paths.

A *phylogenetic tree* is a binary tree with labelled leaves.

A *clade* of a phylogenetic tree is a subset of the tree where an edge is chosen, and everything on one side of the edge is thrown away to create a subtree.

A *cherry* is a clade with 2 leaves. Note that based on topology, cherries using the same leaves will be equivalent.

A *split* of a tree is a partition of the leaf set into two parts.
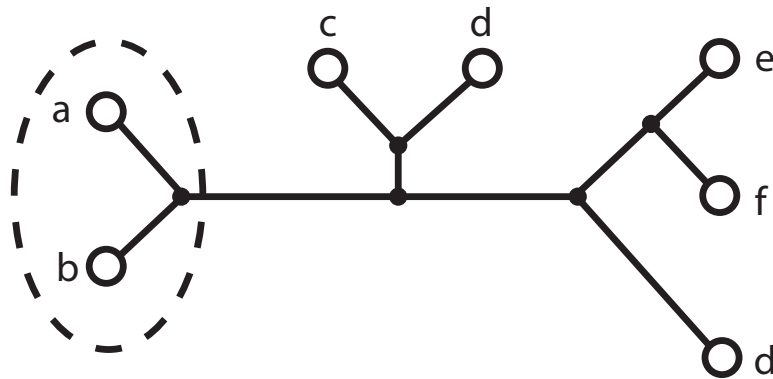

Figure 3.2: This tree is a phylogenetic tree. The circled portion here forms a cherry. We label it a, b. If we consider the partition created by the circle, we have a split. Both the areas inside and outside the circle are clades.

3.2 Combinatorial Polytopes

Besides our geometric definition of a polytope, we can consider polytopes in a more combinatorial sense. While there are objects called abstract polytopes, they include a few objects that do not behave like our geometric polytopes. However, there are several important ideas that we can take away.

One notion is that we can represent a polytope as a poset. To do this, we take our set to be all of the faces of the polytope together with the polytope itself and the empty set, and use containment as the partial order. This will give rise to a lattice since any two faces must have an intersection of a face or nothing at all, and there must be a face or the whole polytope containing both of them.

Another notion is that of a *flag*, or a maximal chain of faces. A flag is a path from the polytope as a whole to the empty set by stepping one face at a time, where the next face must be contained in the previous face. For the poset to be a polytope, every flag must be the same length. The *rank* of a face is defined by using how many steps away from the empty set it is in a flag. If it is m steps away from the empty set, we say the face has rank $m - 1$. This means the empty set itself has rank -1, a vertex has rank 0, an edge rank 1, and so on. We call the maximum rank the *dimension* of the polytope.

3.3 Using the Combinatorial Ideas with Geometry

Recall that we have a face of a polytope when a set of points of the polytope satisfy a linear equality, and the rest of the points of the polytope satisfy a corresponding strict linear inequality. This means that we can represent faces by the linear inequalities which give them in our poset. Since each subface is contained within the face above it, it will satisfy the equality above it.

If we are constructing a flag using only inequalities starting from high rank faces moving to low rank faces, in each new step we only need to make sure that our inequalities exclude points we do not wish to keep from the previous step, meaning we can be more liberal with our choices. This works because each face of a polytope is a polytope itself, so if we show we have a face of a face, that face must also be a face of the whole polytope.

If we can show that a face has rank of $n-1$ where the polytope has dimension n, we have shown that that face is a facet. To do this, we just need to find a chain of $n-1$ inequalities describing faces, where each one of these inequalities is an equality for fewer points then the previous one. This works because it is shown to be a face and is also shown to be inside the other face, so it must be a subface.

CHAPTER IV

THE BALANCED MINIMAL EVOLUTION POLYTOPE

4.1   Introduction to the Balanced Minimal Evolution Polytope

The space given by the convex hull of trees discussed in the introduction is the
Balanced Minimal Evolution Polytope. Loosely speaking, it is a polytope where each
vertex is a binary tree, and if vertices are connected by edges, then the corresponding
trees are similar to one another. This is important since the geometric closeness of
the similar trees minimizes error and allows for a faster construction algorithm.

The solution to the balanced minimal evolution problem is the vertex, or
tree, which has distances between leaves as similar to distances between our species
as possible.

We have a function which sends a pair of two leaves on a binary tree to a
number. This is used in order to create the coordinates which we will take a convex
hull of to create our polytope. The way the function is defined is as follows. $X_{a,b}$,
where a, b are leaves on a tree T, is 2^{n-l-2} where n is the total number of leaves on
the tree and l is the amount of interior nodes on the shortest path between the two
leaves a, b on the tree.

We will now construct the nth BME polytope in the manner that follows [1].

11

For an n-leaf labelled binary tree, we create an ordered list of $\binom{n}{2}$ elements, ordered lexicographically, each element corresponding to a unique pair of leaf labels. This can be written as a function of t like follows: $x(t) =< X_{1,2}, X_{1,3}, X_{1,4}, ..., X_{2,3}, X_{2,4}, ..., X_{n-1,n} >$. We then take the convex hull of the set of outputs of the function. The resulting object is the nth BME polytope.
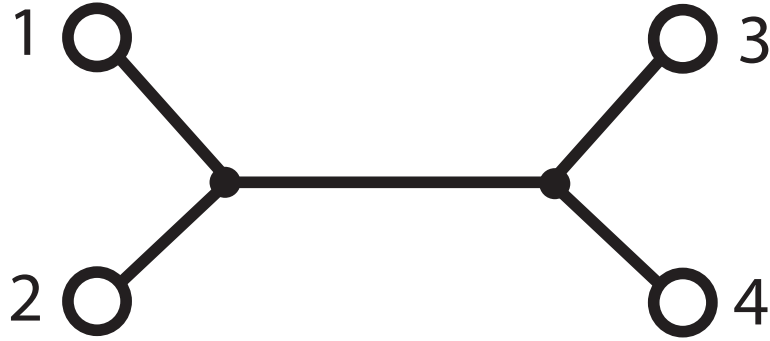

Figure 4.1: We will calculate the x-vector for this tree.

Here our vector is $x(t) =< X_{1,2}, X_{1,3}, X_{1,4}, X_{2,3}, X_{2,4}, X_{3,4} >$, so we need to calculate the outputs of the function X. Here $n = 4$. Since there is 1 internal node between 1 and 2, $X_{1,2} = 2^{4-1-2} = 2$. Similarly, 3 and 4 have 1 internal node between, so $X_{3,4} = 2$ as well. Every other pair of leaves have 2 nodes between them necessarily, so they have value $2^{4-2-2} = 1$. This means our vector is $< 2, 1, 1, 1, 1, 2 >$.

It has been shown [2] that each point used in the construction of the BME polytope is a vertex of the BME polytope.

The dimension of the polytope is $\binom{n}{2} - n$. This is because there are n linear equalities that all points of the polytope obey [1].

The nth BME polytope has $(2n - 5)!!$ vertices since there are that many labelled binary trees on n vertices and as above, each one of these corresponds to a vertex.

The sequence giving the number of facets is not currently known. Lower bounds can be obtained by looking at how different families of facets grow, but this would be so far off that it would not be worthwhile.

4.2   Split Facets

In our work, we have conjectured a new type of facet for the BME polytope. Combinatorially, it is defined by partitioning the leaves of a tree into two sets of size at least 3 each. We then allow the tree to take any topology as long as each set of leaves remains grouped as closely as possible. Note that if one set of leaves is grouped as closely as possible, then the other set must also be. This means that, if we call one set of leaves A, our facet can be given by maximizing $\sum\limits_{i,j \in A} X_{i,j}$. For a given amount of leaves and set A, this sum will remain a constant since this sum does not depend on topology. This gives the inequality for a face. We propose that this set of faces in fact is a set of facets, and demonstrate this in the case where one piece of the partition is of size 3.

13

4.3   A Special Case for Split Facets

Theorem. For any split of pairs size 3 and $m \geq 3$, the trees obeying this split make up the vertices of a facet of the nth BME polytope.

Proof. In the case where $k = 3$, that is one part of the split is size 3, we have a new proof that we have a facet for any m. To do this, we fill in the flag which goes from this facet down to the clade face [2] for a fixed combination of the 3-leaved section of the split.

The first inequality is that of the facet itself, where we simply have a split. If we label the leaves in our k-leaf section a,b,c; we then must have the inequality $X_{a,b} + X_{a,c} + X_{b,c} \leq 2^{n-2}$. This gives a face, which we will show to be the facet, since it requires all of the leaves to be in a clade together to have an equality, and if anything moves out of the clade, it is easy to see that it would become smaller. Let the leaves in the m-leaf section be labeled as $1, 2, ..., m$. We now rely on the fact that our inequalities only need to exclude trees left by the previous equality.

Our next inequality is $3X_{a,1} - X_{b,1} - X_{c,1} + 2X_{a,b} + 2X_{a,c} \leq 3 * 2^{n-3}$. This is intended to include all trees with a in a cherry, and to require the leaf 1 to be near the leaf a when a is not in the cherry.

In the case when a is in the cherry, $X_{b,1}$ or $X_{c,1}$ will be the size of $X_{a,1}$ and the other will be twice its size. So the sum $3X_{a,1} - X_{b,1} - X_{c,1}$ will be 0. Then, $X_{a,b}$ or $X_{a,c}$ must be 2^{n-3} and the other 2^{n-4}. These add to $3 * 2^{n-4}$. So $2X_{a,b} + 2X_{a,c} = 3 * 2^{n-3}$.

When a is not in the cherry, for out inequality to be maximal we must have $X_{a,1} = 2^{n-4}$ and hence $X_{b,1}$ and $X_{c,1}$ as 2^{n-5}. So $3X_{a,1} - X_{b,1} - X_{c,1} = 3*2^{n-4} - 2*2^{n-5} = 2^{n-3}$. Then, since a is near b and c but not in the cherry, we have $2X_{a,b} + 2X_{a,c} = 2*2^{n-3}$. So, the left hand side of our equation is $3*2^{n-3}$ when 1 is close to a, as wanted. If 1 were to be further, it is easy to see the expression would be smaller.

Our next set of steps is dependent upon the size of m. The intent here is to build off of previous steps by forcing specific leaves to be far from the k-leaf cluster in each step. Our inequalities will look like $3X_{a,i} - X_{b,i} - X_{c,i} + \frac{2^{i-1}}{2^{n-4}}(X_{a,b} + X_{a,c}) \geq 3*2^{i-1}$ when $i \geq 3$. When $i = 2$, we require $3X_{a,2} - X_{b,2} - X_{c,2} + \frac{2^{3-1}}{2^{n-4}}(X_{a,b} + X_{a,c}) \geq 3*2^{3-1}$. This is because 2 is in a cherry with 3 so they must satisfy the same inequality, albeit with different coordinates.

This works since $3X_{a,i} - X_{b,i} - X_{c,i}$ is 0 when a is in the cherry, and it is half the size of $\frac{2^{i-1}}{2^{n-4}}(X_{a,b} + X_{a,c})$ when it is not in the cherry. Also, $\frac{2^{i-1}}{2^{n-4}}(X_{a,b} + X_{a,c})$ is $\frac{3}{2}$ the size when a is in the cherry as when a is not in the cherry. So in both cases, when we have what we want, we have equality. If the leaf i moves at all when a is in the cherry, we still have equality. If it moves when a is not in the cherry, $3X_{a,i} - X_{b,i} - X_{c,i}$ will become larger.

After this chain, we have a simple inequality which forces a to be in the cherry. It looks like $2X_{a,b} + 2X_{a,c} \leq 3*2^{n-3}$.

Next, we have $3X_{b,1} - X_{a,1} - X_{c,1} + 2X_{a,b} + 2X_{b,c} \leq 3*2^{n-3}$. This works like the inequality for the face below the facet. This meaning that, it forces 1 to be close

to b when b is not in the cherry, and has no effect on the tree when b is in the cherry.

We then have the same i-indexed chain after it with the roles of a and b reversed, since we are trying to achieve the same result as with a but with b. So, the chain looks like $3X_{b,i} - X_{a,i} - X_{c,i} + \dfrac{2^{i-1}}{2^{n-4}}(X_{b,a} + X_{b,c}) \geq 3 * 2^{i-1}$ when $i \geq 3$ and when $i = 2$, $3X_{b,2} - X_{a,2} - X_{c,2} + \dfrac{2^{3-1}}{2^{n-4}}(X_{b,a} + X_{b,c}) \geq 3 * 2^{3-1}$.

To finish, we do not do the same thing with c, but instead use the fixed clade face of dimension $\binom{m+1}{2} - (m+1)$ [2] where c is not in the cherry. The sum of these parts in our chain is $\binom{n}{2} - n - 1$, proving that the split face is a facet. $\qquad\square$
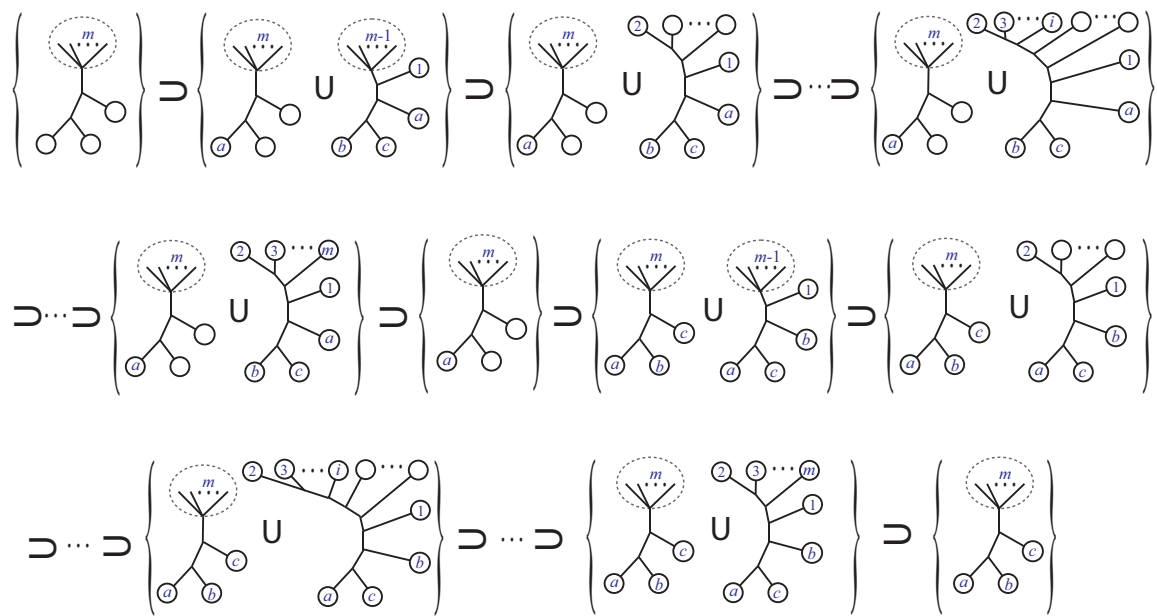
Figure 4.2: The flag used in the proof of this facet

BIBLIOGRAPHY

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

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

[3] Y. Pauplin. Direct calculation of a tree length using a distance matrix. *Journal of Molecular Evolution*, 2000.

[4] A. Rzhetshy and M. Nei. Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Molecular Biology and Evolution*, 1993.