TROPICAL GEOMETRY AND PHYLOGENETIC DIVERSITY

HAN-BOM MOON

1. Introduction - phylogenetic diversity

Since the dawn of tropical geometry, it has been observed that there is an interesting connection between tropical geometry and phylogenetics, which is a branch of biology. An n-species *phylogenetic tree* is an n-leaf metric tree T. Let $\mathcal{T}_n$ be the set of n-species phylogenetic trees. For a fixed metric tree T and a pair (i, j) of leaves, we can measure their *distance* d_{ij}, by taking the sum of all edge lengths on the unique path from i to j. Then we have a distance map

$$\begin{array}{rcl} d : \mathcal{T}_n & \to & \mathbb{R}^{\binom{n}{2}} \\ T & \mapsto & (d_{ij}). \end{array}$$

In computational phylogenetics, a fundamental question is:

Question 1.1. For a given distance information $\mathbf{w} \in \mathbb{R}^{\binom{n}{2}}$, recover the phylogenetic tree T such that $d(T) = \mathbf{w}$.

As trained mathematicians, we should investigate the following questions first: For a given $\mathbf{w}$, is it in $\mathrm{im}\ d$ (so there is a phylogenetic tree)? Is the map d injective (thus, the tree is uniquely recovered)? These natural questions were answered before the tropical geometry era ([Bun71]). Later, in one of the first tropical geometry papers ([SS04]) of Speyer and Sturmfels, it was proved that $\mathrm{im}\ d = \mathrm{Trop}(\mathrm{Gr}(2, n))$, hence tropical geometry naturally appears.

Motivated by the question of finding a more noise-resistant tree reconstruction algorithm, in [PS04], Pachter and Speyer defined the *r-dissimilarity map* (also called *phylogenetic diversity* in biology literature) as the following. Fix an integer $2 \le r \le n-2$. For $T \in \mathcal{T}_n$ and each r-subset $I \subset [n]$, we can define $d_I(T)$ as the total length of the subtree generated by the leaves in I. Then the r-dissimilarity map $d_r : \mathcal{T}_n \to \mathbb{R}^{\binom{n}{r}}$ is defined as $d_r(T) = (d_I) \in \mathbb{R}^{\binom{n}{r}}$. Note that if $r = 2$, $d_r = d$. Pachter and Speyer proved that d_r is injective when $r \le (n+1)/2$, and observed that $\mathrm{im}\ d_r$ seems to be a polyhedral fan in $\mathrm{Trop}(\mathrm{Gr}(r, n))$ (proved later). Thus, the following question is natural:

Question 1.2. Is $\mathrm{im}\ d_r$ a tropical subvariety of $\mathrm{Trop}(\mathrm{Gr}(r, n))$? If so, what is the variety X with $\mathrm{im}\ d_r = \mathrm{Trop}(X)$?

Date: March 26, 2022.

2. Weighted dissimilarity map

In [CGMS21], we investigated Question 1.2. First of all, with a computer-aided computation, it was shown that $\operatorname{im} d_4$ is not a balanced fan when $n = 7$, hence is not a tropical variety ([CGMS21, Theorem 1.1]). However, we were able to modify the definition of the dissimilarity map and obtained a tropical variety as Pachter and Speyer envisioned.

As before, fix $2 \le r \le n-2$. For $T \in \mathcal{T}_n$ and an r-set $I \subset [n]$, now we define

$$d_I^{wt} := \sum_{i<j \in I} d_{ij}.$$

The *weighted r-dissimilarity map* is

$$\begin{array}{rcl} d_r^{wt} : \mathcal{T}_n & \to & \mathbb{R}^{\binom{n}{r}} \\ T & \mapsto & (d_I^{wt}). \end{array}$$

Then $d_2^{wt} = d_2 = d$, $d_3^{wt} = 2d_3$, but d_r^{wt} and d_r are not proportional for $r \ge 4$. We propose that this is a better way to encode the dissimilarity information.

It was shown that d_r^{wt} is a composition of

$$\mathcal{T}_n \stackrel{d_2}{\to} \mathbb{R}^{\binom{n}{2}} \stackrel{L}{\to} \mathbb{R}^{\binom{n}{r}}$$

where L is a full-rank 0-1 matrix. So the injectivity of d_r^{wt} is immediate. The image has to be a tropical variety as well because the matrix L is the tropicalization of a toric morphism $\varphi_r : (\mathbb{C}^*)^{\binom{n}{2}} \to (\mathbb{C}^*)^{\binom{n}{r}}$. Furthermore, φ_r is compatible with a natural rational map $\mathrm{Gr}(2, n) \dashrightarrow \mathrm{Gr}(r, n)$, the so-called *Veronese-Grassmannian map* ([CGMS21, Definition 2.2]).

Theorem 2.1. *For $2 \le r \le n-2$, d_r^{wt} embeds $\mathcal{T}_n$ as a tropical subvariety in $\mathbb{R}^{\binom{n}{r}}$. This tropical variety is the tropicalization of a subvariety $X_{r,n}$ of $\mathrm{Gr}(r, n)$, that is the image of the Veronese-Grassmannian map.*

It is desired to find a finite set of defining equations (*tropical basis*) of $\operatorname{im} d_r^{wt}$. Providing such a set enables us to efficiently check whether a vector $\mathbf{w} = (w_I) \in \mathbb{R}^{\binom{n}{r}}$ is a weighted dissimilarity vector of some phylogenetic tree T.

By the Gelfand-MacPherson correspondence, $X_{r,n}$ is associated with $V_{r-1,n} \subset (\mathbb{P}^{r-1})^n$, the space parametrizing n-point configurations lying on a rational normal curve in $\mathbb{P}^{r-1}$. Then the set of $(\mathbb{C}^*)^n$-invariant polynomials defining $X_{r,n}$ is identified with the set of SL_r-invariant polynomials defining $V_{r-1,n}$. In [CGMS18], an inductive algorithm generating such polynomials is described. It is based on the reduction via the Gale duality

$$V_{r-1,n}/\!/\mathrm{SL}_r \cong V_{n-r-1,n}/\!/\mathrm{SL}_{n-r},$$

and the pull-back of equations via $V_{r-1,n} \to V_{r-1,n-1}$. In [CGMS21], it was shown that two collections of equations, one coming from the tropicalization of three-term Plücker relations for $\mathrm{Trop}(\mathrm{Gr}(r, n))$ and the other coming from the tropicalization of the above equations for $V_{r-1,n}$, are sufficient to cut out $\operatorname{im} d_r^{wt}$.

Theorem 2.2. *A vector* $\mathbf{w} = (w_I) \in \mathbb{R}^{\binom{n}{r}}$ *is in* $\operatorname{im} d_r^{wt}$ *if and only if:*

(1) *for each 4-set* $\{i, j, k, \ell\} \subset [n]$*, there exists* $I \subset [n] \setminus \{i, j, k, \ell\}$ *of size* $r - 2$ *such that the maximum of the three terms below is achieved twice:*

$$w_{ijI} + w_{k\ell I},\ w_{ikI} + w_{j\ell I},\ w_{i\ell I} + w_{jkI};$$

(2) *for each* $I \in \binom{[n]}{6}$*,* $J \in \binom{[n]\setminus I}{r-3}$*, and for each cube* C *on* I *(see* [CGMS21, Section 5.2] *for the notation) with corresponding bipartition* B, W *we have*

$$\sum_{K \in B} w_{J \sqcup K} = \sum_{K \in W} w_{J \sqcup K}.$$

When $r = 2$, the equations in (1) are precisely the 'four-point conditions' of [Bun71].

3. Further questions

Here we leave three open questions.

Question 3.1. Find a higher-dimensional analogue of the main theorems. For instance, does the space of contractible metrized simplicial complexes admit a tropical variety structure?

Question 3.2. A natural generalization of $V_{d,n}$ is the space of point configurations on positive genus curves. As a first example, what is the tropicalization of the space of point configurations on elliptic curves in $\mathbb{P}^2$?

Question 3.3. The *neighbor-joining algorithm* ([SN87]) is a revolutionary algorithm for reconstructing the phylogenetic tree from a given tree metric. Find a similar tree-reconstruction algorithm based on the weighted dissimilarity map.

References

[Bun71] P. Buneman, The recovery of trees from measures of dissimilarity. In: *Mathematics in the Archaeological and Historical Sciences* (ed. F. Hodson, D.Kendall, and P. Tautu), Edinburgh University Press, Edinburgh (1971), 387–395. 1

[CGMS18] A. Caminata, N. Giansiracusa, H.-B. Moon, and L. Schaffler, Equations for point configurations to lie on a rational normal curve. *Adv. Math.* 340 (2018), 653–683. 2

[CGMS21] A. Caminata, N. Giansiracusa, H.-B. Moon, and L. Schaffler, Point configurations, phylogenetic trees, and dissimilarity vectors. *Proc. Natl. Acad. Sci. USA*, 2021, 118 (12). 2, 3

[PS04] L. Pachter and D. Speyer, Reconstructing trees from subtree weights. *Appl. Math. Lett.* 17 (2004), no. 6, 615–621. 1

[SN87] N. Saitou and M. Nei, The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, Volume 4, Issue 4, Jul 1987, Pages 406–425. 3

[SS04] D. Speyer and B. Sturmfels, The tropical Grassmannian. *Adv. Geom.* 4 (2004), no. 3, 389–411. 1

Department of Mathematics, Fordham University, New York, NY 10023

Email address: hmoon8@fordham.edu