QUANTIFYING BARLEY MORPHOLOGY USING THE EULER CHARACTERISTIC TRANSFORM

Anonymous authors
Paper under double-blind review

ABSTRACT

Shape plays a fundamental role in biology. Throughout history, the constant observation and documentation of the shape of multiple organisms has been key to further biological understanding of organism and tissue behavior, function, and origins. To extract, compare, and analyze such data diversity we must wield a quantifiable, robust, and concise method. We thus turn to Topological Data Analysis (TDA) and the Euler Characteristic Transform (ECT). As a study case, we quantify the morphology of X-ray CT scans of barley spikes and seeds using both traditional and topological shape descriptors based. We then successfully train a support vector machine to distinguish and classify 28 different parental genotypes of barley based solely on the 3D shape of their grains. We observe that combining both traditional and topological descriptors produces considerably better classification results compared to the use of exclusively traditional descriptors. This improvement suggests that TDA is thus a powerful complement to describe comprehensively a multitude of shape nuances which are otherwise not picked up by traditional morphometrics methods.

1 INTRODUCTION

Shape is data and data is shape. Biologists are accustomed to thinking about how the shape of biomolecules, cells, tissues, and organisms arise from the effects of genetics, development, and the environment. Traditionally, biologists use morphometrics to compare and describe shapes. The shape of leaves and fruits is quantified based on homologous landmarks—similar features due to shared ancestry from a common ancestor—or harmonic series from a Fourier decomposition of their closed contour. While these methods are useful for comparing many shapes in nature, they cannot always be used: there may not be homologous points between samples or a harmonic decomposition of a shape is not appropriate. Topological data analysis (TDA) offers a more comprehensive, versatile way to quantify plant morphology. In particular, Euler characteristic curves (ECC) serve as a succinct, computationally feasible topological signature that allows downstream statistical analyses (Turner et al., 2014). For example, Li et al. (2018) computed a morphospace for all leaves and used ECCs to predict plant family and location, to later determine the genetic basis of leaf shape in apple (Migicovsky et al., 2018), tomato (Li et al., 2018), and cranberry (Diaz-Gárcia et al., 2018). ECCs are sensitive enough to detect both complex and subtle effects of rootstock and climate on grapevine leaf shape (Migicovsky et al., 2019), the shape of spikelets—arrangements of grass flowers—and their hairiness (McAllister et al., 2019), and patterns of vegetation from satellite imagery (Mander et al., 2017). Here, we show the use of the Euler characteristic to comprehensively describe the shape of barley seeds as a proof of concept.

2 METHODS

Consider a cubical complex X of dimension d. For a fixed direction $\nu \in S^{d-1}$, and a height value $h \in \mathbb{R}$, we define

$$X(\nu)_h = \{\Delta \in X \,:\, \langle x, \nu \rangle \leq h \text{ for all } x \in \Delta\},$$

to be the subcomplex containing all cubical cells below height h in the direction ν. The Euler characteristic at height h is $\chi(X(\nu)_h)$, the alternating sum of counts of cells in the subcomplex $X(\nu)_h$. The Euler Characteristic Curve (ECC) of direction ν is defined as $\{\chi(X(\nu)_h)\}_{h\in\mathbb{R}}$, exemplified in

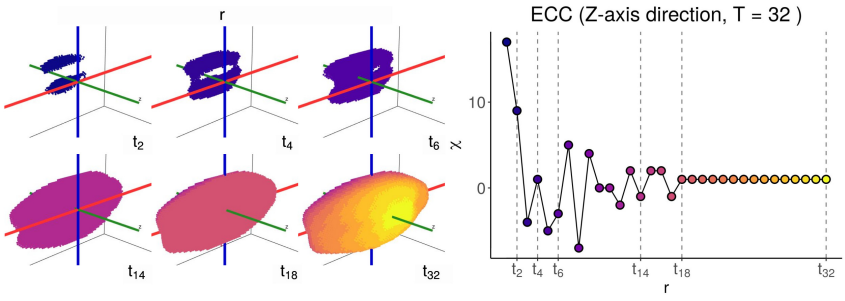


Figure 1: Filtration of a barley seed along the z-axis with 32 thresholds and its corresponding Euler Characteristic Curve.

Figure 1. The Euler Characteristic Transform (ECT) is defined as the collection of all ECCs corresponding to all possible directions. To be more precise, the ECT of complex X is defined as the function

$$\begin{aligned} ECT(X) : S^{d-1} &\to \mathbb{Z}^{\mathbb{R}} \\ \nu &\mapsto \{\chi(X(\nu)_h)\}_{h\in\mathbb{R}}. \end{aligned}$$

We are studying the morphology of barley seeds and barley spikes—the branching inflorescence. We focus on a collection of 28 different parental barley genotypes from diverse regions across the Eurasian continent. Using X-ray CT—computed tomography—scanning technology, we have created voxel-based 3D reconstructions of over 875 spikes, from which we have isolated more than 3100 parental seeds. Since the seeds are oblong in shape, we aligned them according to their three main principal components.

On one hand, we computed 11 traditional quantifiable shape descriptors for each seed, such as length, height, width, volume, and surface area. On the other hand, we computed topological shape descriptors using the ECT. For topological purposes, we treated each voxel-based image as a dual cubical complex where each nonzero voxel is treated as a vertex (Wagner et al., 2012).

We favor the use of the ECT for two reasons. First, the ECT is computationally inexpensive, since it is based on successive alternating sums of counts of cells. Following a strategy similar to the one outlined by Richardson and Werman (2014), computing a single ECC is an $O(N)$ operation with respect to the number of voxels in the image. This inexpensiveness is especially relevant as we deal with thousands of extremely high-resolution 3D images. Second, Turner et al. (2014) proved that the ECT effectively summarizes all the morphological features of any 3D complex as it encodes sufficient information to reconstruct the initial complex. Such results were later extended to the n-dimensional case independently by Curry et al. (2018) and Ghrist et al. (2018). Later, Curry et al. (2018) proved a finite bound on the number of necessary directions for general 3D shapes, although the idea of efficiently reconstructing an arbitrary 3D object solely from its ECT remains elusive (Betthauser, 2018; Fasy et al., 2019; Micka, 2020).

In our case, we used 158 different directions with 8 uniformly spaced thresholds. We emphasized directions toward the seed's cleft, which correspond to directions close to both north and south poles as shown in Figure 2a. This yielded a 1264-dimensional vector for every seed. Due to statistical problems associated with high dimensionality, we reduced all ECT vectors to just 12 dimensions using a kernel principal component analysis (KPCA) with a Laplace kernel (Schölkopf et al., 1998).

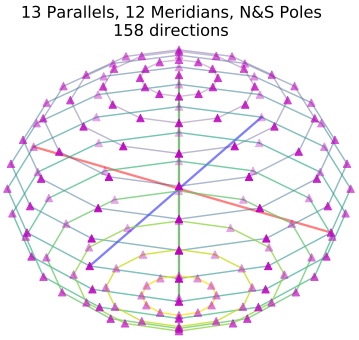
We then sought to test the descriptiveness of both traditional and topological measures. To this end, we trained three non-linear support vector machines (SVM) (Burges, 1998) to characterize and predict the seeds from the 28 different parental genotypes based on three different collection of descriptors: traditional, topological, and combining both traditional and topological descriptors. In every case, the descriptors were centered and scaled to variance 1 prior to classification. Given that SVM is a supervised learning method, we partitioned our data into training and testing sets. In our case, we randomly sampled 80% of the seeds from every founder as our training data set. The remaining 20% was used to test the accuracy of our prediction model. We repeated this SVM setup 100 times and considered the average accuracy and confusion matrices as final results. The 25th and 75th quantiles of classification accuracy are reported in Table 1. Average accuracy values for all barley varieties are shown in Figure 3.

Table 1: SVM classification accuracy of barley seeds from 28 different founding lines after 100 randomized training and testing sets. The ECT was computed with 158 directions (as in Figure 2a) and 8 thresholds.

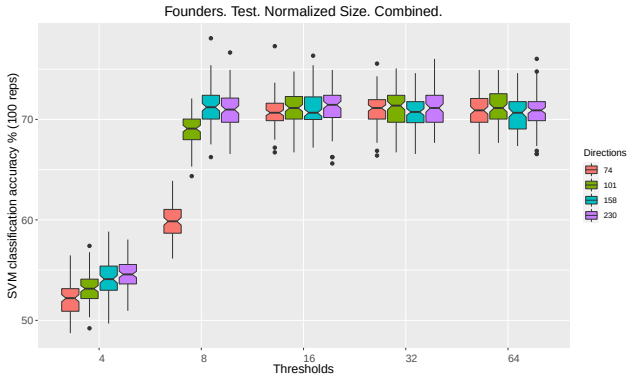
Shape descriptors	No. of descriptors	Classification accuracy (Q25—Q75)
Traditional	11	54.7%—57.1%
Topological (ECT + KPCA)	12	53.9%—56.9%
Combined (Traditional + Topological)	23	70.0%—72.4%

Table 2: Quade post-hoc p-values (with Bonferroni correction) to determine if different descriptors produce statistically different SVM results

Assuming t distribution			Assuming normal distribution		
	Traditional	Topological		Traditonal	Topological
Topological	8.6×10^{-3}	$*$	Topological	6.7×10^{-5}	$*$
Combined	$< 2 \times 10^{-16}$	$< 2 \times 10^{-16}$	Combined	$< 2 \times 10^{-16}$	$< 2 \times 10^{-16}$



(a) Directions chosen to compute the ECT



(b) Overall SVM classification accuracy when using combined shape descriptors. The topological descriptors were computed with different threshold and direction parameters

Figure 2: We used 158 directions and 8 thresholds each to compute the ECTs. We chose this parameters, as increasing either the number of directions or thresholds didn't improve the SVM classification results.

Carrying out a Friedman test (Friedman, 1937) to determine if there is a statistical difference between the three SVM classifiers, we obtain a p-value less than 2×10^{-16}, which suggests significant difference. Since we are comparing only three classifiers, we can rely better on a Quade test (Quade, 1979) as suggested in Conover (1998). This produces a p-value smaller than 2×10^{-16} as well. The significance prompts a pairwise comparison test. The p-values are reported in Table 2. Recall that the results presented on Tables 1 and 2 and Figure 3 are based on an ECT computed with 158 directions and 8 thresholds. We chose these parameters on the observation that increasing the number of thresholds or directions did not improve overall classification results as shown in Figure 2b.

3 Results and Conclusions

The Euler characteristic is a simple yet powerful way to reveal features not readily visible to the naked eye. The small p-values for the Quade test seem to confirm that the SVM classifier with combined descriptors is statistically distinct from the classifiers relying exclusively in either traditional or topological descriptors. These small p-values remained small as we evaluated other post-hoc tests, such as Nemenyi and Conover with different p-value corrections. Nonetheless, we are aware

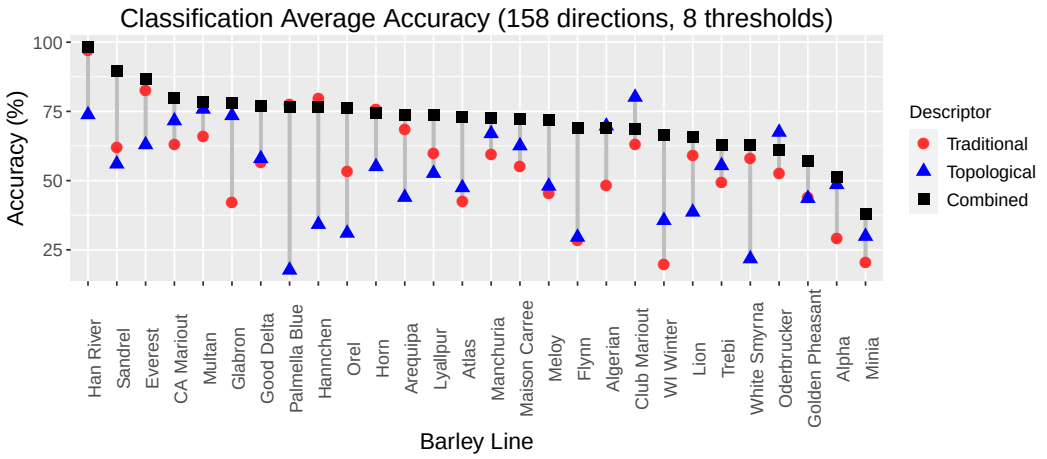


Figure 3: Average classification accuracy for different barley genotypes when using combined shape descriptors. The topological descriptors were computed with 158 directions and 8 thresholds each.

that a more careful statistical analysis is necessary, as the combined SVM is naturally dependent with both traditionally and topologically based SVM.

Even though the use of exclusively traditional or topological descriptors produces the same overall classification accuracy as seen in Table 1, observe that certain barley varieties are better distinguishable with the topological lens but not with traditional measures, and vice-versa. For instance, Glabron and Algerian report considerably higher classification accuracies whenever using topological information compared to using only traditional measures. Moreover, some lines such as Club Mariout are better characterized using exclusively topological features, as combining traditional measures just muddles classification results. On the other hand, our topological descriptors perform poorly whenever we try to distinguish lines such as Palmella Blue and Hannchen, as these lines seem better characterized by traditional measures. Finally, some lines like Wisconsin Winter reported poor classification results whenever we limited ourselves to just topological or traditional measures; however, our classification accuracy improved dramatically as we combined both descriptors.

Natural variation in barley, like all crops, encompasses differences in yield and adaptation to diverse climates and terrains. Understanding how differences in morphology affect these traits is vital to improve barley through breeding. TDA combined with X-ray CT scans offers a novel insight into the plant form and its evolution. As a long term plan, we will compare the topological descriptors to available genetic information of each barley sample. This analysis can further our understanding of the relationship between phenotype and genotype.

REFERENCES

Betthauser LM (2018). *Topological reconstruction of grayscale images*. Ph. D. thesis, University of Florida, Gainesville, Florida.

Burges CJ (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* **2**(2), 121–167. doi:10.1023/A:1009715923555.

Conover WJ (1998). *Practical Nonparametric Statistics* (3rd ed.). Wiley Series in Probability and Statistics. Wiley.

Curry J, Mukherjee S, Turner K (2018). How many directions determine a shape and other sufficiency results for two topological transforms. arXiv:1805.09782.

Diaz-Gárcia L, Covarrubias-Pazaran G, Schlautman B, Grygleski E, Zalapa J (2018). Image-based phenotyping for identification of QTL determining fruit shape and size in american cranberry (*Vaccinium macrocarpon L.*). *PeerJ* **6**(e5461). doi:10.7717/peerj.5461.

Fasy BT, Micka S, Millman DL, Schenfisch A, Williams L (2019). The first algorithm for reconstructing simplicial complexes of arbitrary dimension from persistence diagrams. arXiv:1912.12759.

Friedman M (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association* **32**(200), 675–701. doi:10.1080/01621459.1937.10503522.

Ghrist R, Levanger R, Mai H (2018). Persistent homology and Euler integral transforms. *Journal of Applied and Computational Topology* **2**(1), 55–60. doi:10.1007/s41468-018-0017-1.

Li M, An H, Angelovici R, Bagaza C, Batushansky A, Clark L, Coneva V, Donoghue MJ, Edwards E, Fajardo D et al. (2018). Topological data analysis as a morphometric method: Using persistent homology to demarcate a leaf morphospace. *Frontiers in Plant Science* **9**, 553. doi:10.3389/fpls.2018.00553.

Li M, Frank MH, Coneva V, Mio W, Chitwood DH, Topp CN (2018). The persistent homology mathematical framework provides enhanced genotype-to-phenotype associations for plant morphology. *Plant Physiology* **177**(4), 1382–1395. doi:10.1104/pp.18.00104.

Mander L, Dekker SC, Li M, Mio W, Punyasena SW, Lenton TM (2017). A morphometric analysis of vegetation patterns in dryland ecosystems. *Royal Society Open Science* **4**(2), 160443. doi:10.1098/rsos.160443.

McAllister CA, McKain MR, Li M, Bookout B, Kellogg EA (2019). Specimen-based analysis of morphology and the environment in ecologically dominant grasses: the power of the herbarium. *Philosophical Transactions of the Royal Society B: Biological Sciences* **374**(1763), 20170403. doi:10.1098/rstb.2017.0403.

Micka SA (2020). *Searching and Reconstruction: Algorithms with Topological Descriptors*. Ph. D. thesis, Montana State University, Bozeman, Montana.

Migicovsky Z, Harris ZN, Klein LL, Li M, McDermaid A, Chitwood DH, Fennell A, Kovacs LG, Kwasniewski M, Londo JP et al. (2019). Rootstock effects on scion phenotypes in a *Chambourcin* experimental vineyard. *Horticulture Research* **6**(64). doi:10.1038/s41438-019-0146-2.

Migicovsky Z, Li M, Chitwood DH, Myles S (2018). Morphometrics reveals complex and heritable apple leaf shapes. *Frontiers in Plant Science* **8**, 2185. doi:10.3389/fpls.2017.02185.

Quade D (1979). Using weighted rankings in the analysis of complete blocks with additive block effects. *Journal of the American Statistical Association* **74**(367), 680–683. doi:10.1080/01621459.1979.10481670.

Richardson E, Werman M (2014). Efficient classification using the Euler characteristic. *Pattern Recognition Letters* **49**, 99 – 106. doi:10.1016/j.patrec.2014.07.001.

Schölkopf B, Smola A, Müller KR (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation* **10**(5), 1299–1319. doi:10.1162/089976698300017467.

Turner K, Mukherjee S, Boyer DM (2014). Persistent homology transform for modeling shapes and surfaces. *Information and Inference* **3**(4), 310–344. doi:10.1093/imaiai/iau011.

Wagner H, Chen C, Vuçini E (2012). Efficient computation of persistent homology for cubical data. In Peikert R, Hauser H, Carr H, Fuchs R (Eds.), *Topological Methods in Data Analysis and Visualization II: Theory, Algorithms, and Applications*, pp. 91–106. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-23175-9_7.