Quantifying barley morphology using the Euler characteristic transform

Erik J. Amézquita[*]
Department of Computational Mathematics, Science & Engineering
Michigan State University

Michelle Quigley
Department of Horticulture
Michigan State University

Tim Ophelders
Department of Mathematics and Computer Science
TU Eindhoven

Daniel H. Chitwood
Department of Horticulture
Michigan State University

Elizabeth Munch
Department of Computational Mathematics, Science & Engineering Dept.
Michigan State University

Jacob Landis
School of Integrative Plant Science
Cornell University

Daniel Koenig
Department of Botany & Plant Sciences
University of California—Riverside

Abstract

Shape is foundational to biology. Observing and documenting shape has fueled biological understanding, and from this perspective, it is also a type of data. The vision of topological data analysis, that data is shape and shape is data, will be relevant as biology transitions into a data-driven era where meaningful interpretation of large data sets is a limiting factor. We focus first on quantifying the morphology of barley spikes and seeds using topological descriptors based on the Euler characteristic. We then successfully train a support vector machine to classify 28 different varieties of barley based solely on the shape of their grains.

---

[*] Contact email: `amezqui3@msu.edu`

Topological Data Analysis and Beyond Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

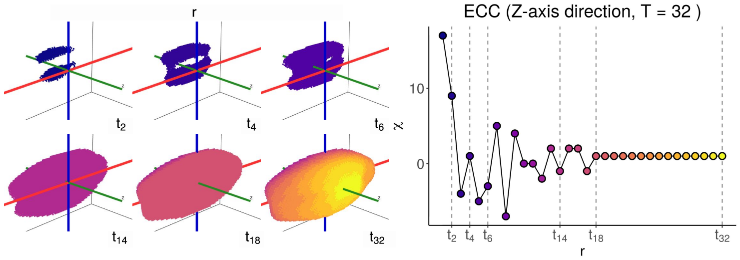

Figure 1: Filtration of a barley seed along the z-axis and its corresponding Euler Characteristic Curve.

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 can not 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 [14] serve as a succinct, computationally feasible topological signature that allows downstream statistical analyses. For example, Li et al. [8] computed a morphospace for all leaves and used ECCs to predict plant family and location. Others have used the same filter and ECCs to determine the genetic basis of leaf shape in apple [10] and tomato [9] as well as the genetic basis of cranberry shape [4].

2 Methods

We are studying the morphology of barley seed and barley spikes (the branching inflorescence). We focus on a collection of 28 different parental barley genotypes from diverse regions. Using X-ray CT 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 quantifiable traditional shape descriptors for each seed, such as length, height, width, volume, and surface area. On the other hand, we also sought to compute topological shape descriptors. Given the large number of seeds and voxels per seed, we use the Euler Characteristic Transform as in [14] to quantify the morphology of each barley spike.

Consider each voxel-based image as a cubical complex X of dimension $d = 3$ as in [12], with each voxel regarded as a vertex. For a fixed direction $\nu \in S^{d-1}$, and a height value h, we define

$$X(\nu)_h = \{\Delta \in X \,:\, \langle x, \nu\rangle \leq h \text{ for all } x \in \Delta\}, \tag{1}$$

to be the subcomplex containing all cubical simplices below height h in the direction ν. The Euler characteristic at height h is $\chi(X(\nu)_h) = V - E + S - C$ where V, E, S, C are the number of vertices, edges, squares and cubes in $X(\nu)_h$ respectively. The Euler characteristic curve (ECC) of direction ν is defined as $\{\chi(X(\nu)_h)\}_{h\in\mathbb{R}}$, as exemplified in Figure 1. Turner et al. [14] proved that the collection of all ECCs corresponding to all possible directions effectively summarizes all the morphological information of 3D shapes in general. Moreover, with such collection we would be able to reconstruct the original object. A finite bound on the number of necessary directions for general 3D shapes has been proven [3], although the idea of efficiently reconstructing arbitrary 3D objects solely from ECCs [5] remains elusive. In our case, we used 74 different directions with 32 uniformly spaced thresholds. This yielded a 2368-dimensional vector for every seed. Due to statistical problems associated with high-dimensional vectors [7], we reduced the ECT vector to just 12 dimensions using kernel principal component analysis (KPCA) [13].

We then sought to classify 28 barley varieties based on their seed morphology using a nonlinear support vector machine (SVM) [1]. We considered 3 different sets of shape descriptors for the barley seeds: traditional descriptors, reduced-dimension ECTs, and a combination of both traditional and topological information. For each set, a training set was formed by randomly selecting 80% of seeds

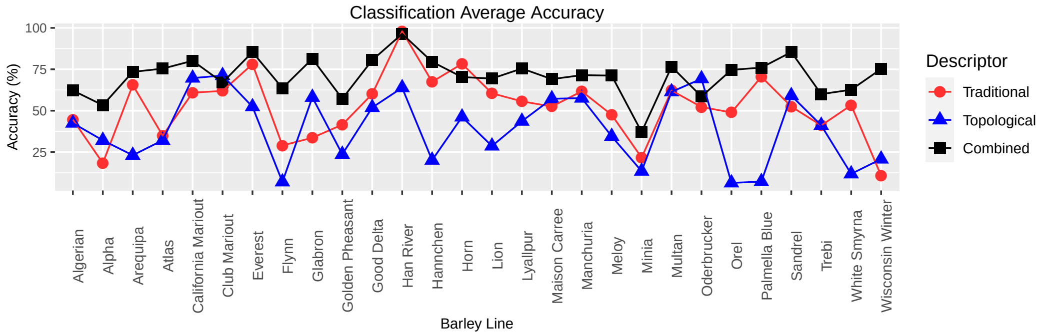

Figure 2: Classification accuracy for all barley varieties using different shape descriptors.

corresponding to every barley variety. The remaining 20% was kept for testing. This sampling of training and testing was repeated 100 times. The 25th and 75th quantiles of classification accuracy are reported in Table 1. Exact accuracy values for all varieties are shown in Figure 2.

Table 1: SVM classification accuracy of barley seeds from 28 different founding lines

Shape descriptors	No. of descriptors	Classification accuracy
Traditional	11	51.9%—54.2%
Topological (ECT + KPCA)	12	43.2%—45.7%
Combined (Traditional + Topological)	23	69.2%—71.9%

Carrying out a Friedman test [6] to determine if there's a statistical difference between the three SVM classifiers, we obtain a p-value of 8.16×10^{-8}, which suggests significance. Since we are comparing only three classifiers, we can rely better on a Quade post-hoc pairwise test [11] as suggested in [2]. The p-values are reported in Table 2.

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	0.031	$*$	Topological	0.0047	$*$
Combined	1.05×10^{-5}	5.4×10^{-10}	Combined	2.2×10^{-9}	$< 2 \times 10^{-16}$

3 Results and Conclusions

The ECT by itself performs overall worse than traditional shape descriptors. However, when both topological and traditional information is combined, we observe a substantial classification improvement compared to using traditional information alone. This is even more striking if we consider that we reduced aggressively the dimension of the ECTs. We also highlight that for some specific barley varieties, the ECT descriptor performed considerably better than its traditional counterpart. This suggests that ECT is sensitive to certain physical traits which are not obvious to the naked eye.

The small p-values for Friedman and Quade tests seem to confirm that the SVM classifiers with different shape descriptors are effectively statistically distinct. 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 that a more careful statistical analysis is necessary, as the combined SVM is naturally dependent with both traditionally and topologically based SVM.

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

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

[2] Conover WJ. *Practical Nonparametric Statistics*. Wiley Series in Probability and Statistics. Wiley, 3rd edition, 1998.

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

[4] Diaz-Gárcia L, et al. Image-based phenotyping for identification of QTL determining fruit shape and size in american cranberry (*Vaccinium macrocarpon L.*). *PeerJ*, 6(e5461), 2018. `doi:10.7717/peerj.5461`.

[5] Fasy BT, et al. The first algorithm for reconstructing simplicial complexes of arbitrary dimension from persistence diagrams. 2019. `arXiv:1912.12759`.

[6] Friedman M. 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, 1937. `doi:10.1080/01621459.1937.10503522`.

[7] Köppen M. The curse of dimensionality. In *5th Online World Conference on Soft Computing in Industrial Applications (WSC5)*, volume 1, pages 4–8, Sept. 2000.

[8] Li M, et al. Topological data analysis as a morphometric method: Using persistent homology to demarcate a leaf morphospace. *Frontiers in Plant Science*, 9:553, 2018. `doi:10.3389/fpls.2018.00553`.

[9] Li M, et al. The persistent homology mathematical framework provides enhanced genotype-to-phenotype associations for plant morphology. *Plant Physiology*, 177(4):1382–1395, 2018. `doi:10.1104/pp.18.00104`.

[10] Migicovsky Z, et al. Morphometrics reveals complex and heritable apple leaf shapes. *Frontiers in Plant Science*, 8:2185, 2018. `doi:10.3389/fpls.2017.02185`.

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

[12] Robins V, Wood PJ, Sheppard AP. Theory and algorithms for constructing discrete Morse complexes from grayscale digital images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1646–1658, Aug. 2011. `doi:10.1109/TPAMI.2011.95`.

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

[14] Turner K, Mukherjee S, Boyer DM. Persistent homology transform for modeling shapes and surfaces. *Information and Inference*, 3(4):310–344, 12 2014. `doi:10.1093/imaiai/iau011`.