Quantifying barley morphology using the Euler Characteristic Transform

Erik Amezquita[1], Michelle Quigley[1], Tim Ophelders[2], Jacob B Landis[3], Daniel Koenig[4], Elizabeth Munch[1], and Daniel H. Chitwood[1]

[1]Michigan State University
[2]TU Eindhoven
[3]Cornell University
[4]University of California-Riverside

November 21, 2022

Abstract

Observing and documenting shape has fueled biological understanding as the shape of biomolecules, cells, tissues, and organisms arise from the effects of genetics, development, and the environment. The vision of Topological Data Analysis (TDA), that data is shape and shape is data, will be relevant as biology transitions into a data-driven era where meaningful interpretation of large datasets is a limiting factor. We focus first on quantifying the morphology of X-ray CT scans of barley spikes and seeds using topological descriptors based on the Euler Characteristic Transform. We then successfully train a support vector machine to distinguish and classify 28 different varieties of barley based solely on the 3D shape of their grains. This shape characterization will allow us later to link genotype with phenotype, furthering our understanding on how the physical shape is genetically coded in DNA.

Quantifying barley morphology using the Euler Characteristic Transform

Erik J. Amézquita[1], Michelle Y. Quigley[2], Tim Ophelders[4], Jacob B. Landis[5,6], Daniel Koenig[6], Daniel H. Chitwood[1,2], and Elizabeth Munch[1,3]

[1]Dept. of Computational Mathematics, Science, & Engineering, Michigan State University
[2]Dept. of Horticulture, Michigan State University
[3]Dept. of Mathematics, Michigan State University
[4]Dept. of Mathematics and Computer Science, TU Eindhoven
[5]School of Integrative Plant Science, Cornell University
[6]Dept. of Botany & Plant Sciences, University of California—Riverside

ABSTRACT

Observing and documenting shape has fueled biological understanding as the shape of biomolecules, cells, tissues, and organisms arise from the effects of genetics, development, and the environment. The vision of Topological Data Analysis (TDA), that data is shape and shape is data, will be relevant as biology transitions into a data-driven era where meaningful interpretation of large datasets is a limiting factor. We focus first on quantifying the morphology of X-ray CT scans of barley spikes and seeds using topological descriptors based on the Euler Characteristic Transform. We then successfully train a support vector machine to distinguish and classify 28 different varieties of barley based solely on the 3D shape of their grains. This shape characterization will allow us later to link genotype with phenotype, furthering our understanding on how the physical shape is genetically coded in DNA.

Keywords: Topological Data Analysis; Euler Characteristic Transform; barley inflorescense; mathematical biology; data science; shape

1. INTRODUCTION

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[1]—similar features due to shared ancestry from a common ancestor—or harmonic series from a Fourier decomposition of their closed contour.[2] 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 available in 3D voxel-based images. Topological Data Analysis (TDA) is a set of tools that arise from the perspective that all data has shape and that shape is data[3–5] and offers a more comprehensive, versatile way to quantify plant morphology. In particular, the Euler characteristic, represented by the Greek letter χ, is a topological invariant; that is, it will remain unchanged under any smooth transformation applied to our shape. It serves as a succinct, computationally feasible topological signature that allows downstream statistical analyses.[6]

We can extract information out of our image if we think of it as a dynamic object that grows in number of voxels across time. As our image grows, we may observe significant changes in χ. The changes in χ can be thought as a topological signature of the shape, referred to as an *Euler characteristic curve (ECC)*. The growth of the complex is defined by a *directional filter function* which assigns to each voxel its height as if measured from a fixed direction (Fig. 1F). To get a better sense of how the Euler characteristic changes overall, we can compute several ECCs corresponding to different directional filters. Each filter produces an individual ECC, which we later concatenate into a unique large signal known as the *Euler Characteristic Transform (ECT)*.

Further information and correspondence:
E. A.: E-mail: amezqui3@msu.edu, Address: 428 S Shaw Ln, 1515 Engineering Building, East Lansing, MI 48824

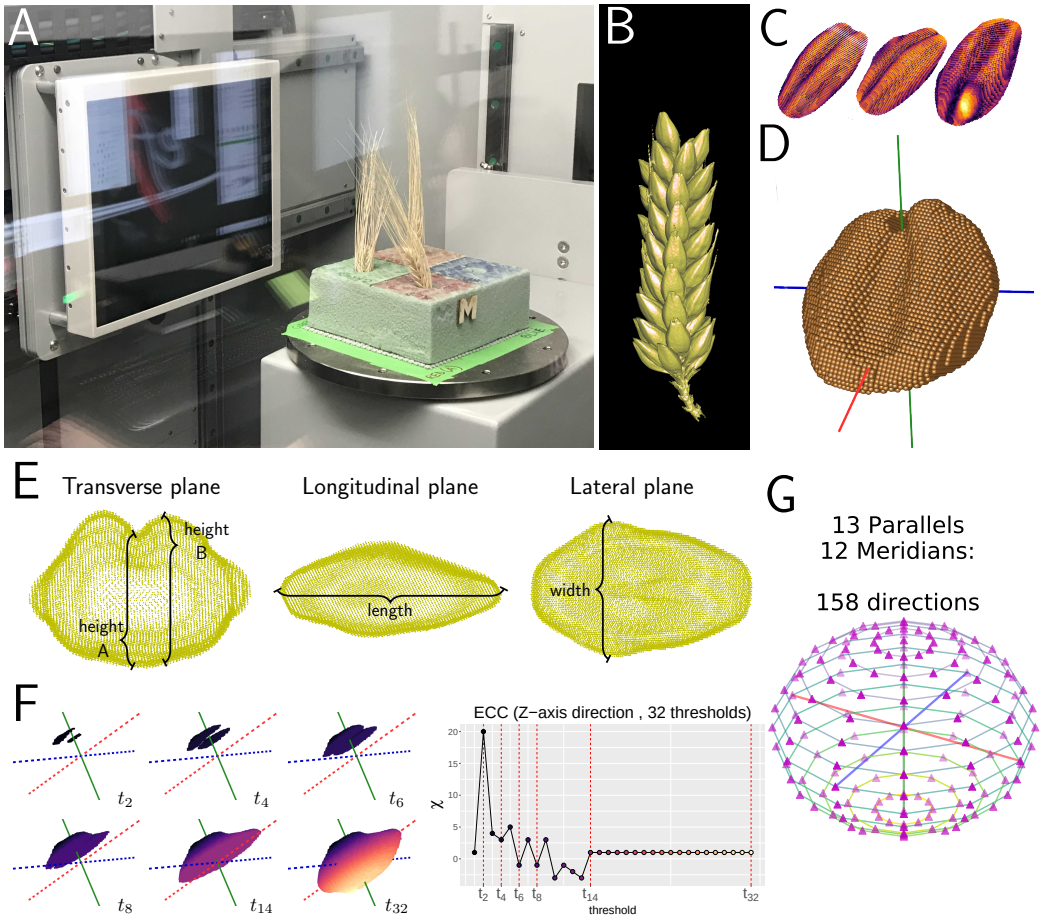

Figure 1. **Barley image processing and measurement.** **(A)** The morphology measurements were extracted from 3D X-ray CT scans of the barley panicles. **(B)** Densities normalized, air and other debris removed, and awns pruned. **(C)** An extra digital step segmented the individual seeds —embryo and endosperm— for each barley spike. **(D)** The seeds were aligned according to their principal components, which allowed us to **(E)** measure a number of traditional shape descriptors. **(F)** Example of an ECC as we filter the barley seed across the adaxial-abaxial axis —depicted as a solid, green line— through 32 equispaced thresholds. **(F)** In total, 158 directions were considered to compute the ECT.

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, it can be computed in linear time with respect to the number of voxels.[7] Second, the ECT effectively summarizes all the morphological features of any 3D complex as it encodes sufficient information to reconstruct the initial complex,[6] a result later extended to the n-dimensional case.[8]

In practice, ECCs has been used to determine a morphospace for all leaves to then predict plant family and location.[9] Further analysis has determined the genetic basis of 2D leaf shape in apple,[10] tomato,[11] and cranberry.[12] Here, we show the use of the Euler Characteristic to comprehensively describe the shape of 3D voxel-based X-ray CT scans of barley seeds as a proof of concept.

2. METHODS

We focus on the shape of barley from a collection of 28 different accesions from diverse regions across the Eurasian continent.[13–15] Seeds from each accession were stratified at 4C on wet paper towels for a week, and germinated on the bench at room temperature. Four day old seedlings were transferred into pots in triplicate and arranged in a completely randomized design in a greenhouse. After the plants reached maturity and dried, a single spike was collected from each replicate for scanning (Fig. 1A). The scans and 3D reconstruction of the spikes were produced using the North Star Imaging X3000 system and the included efX software, obtaining a final voxel size of 127 microns. The densities were standardized across scans, the air and debris were thresholded out, and awns digitally pruned (Fig. 1B). The seed coat of the caryopses was digitally removed, leaving only the embryo and

Table 1. SVM classification accuracy of barley seeds from 28 different accessions after 100 randomized training and testing sets. Classification scores were computed for each accession; the weighted average for each score was taken afterwards, depending on the number of test seeds used. The use of topological descriptors outperforms the use of exclusively traditional ones.

Shape descriptors	Dimension reduction	No. of dims	Scores (weighted average ± standard deviation) Precision	Recall	F_1
Traditional	*	11	0.58 ± 0.050	0.58 ± 0.016	0.57 ± 0.016
Topological	UMAP	12	0.75 ± 0.047	0.75 ± 0.016	0.74 ± 0.016
Combined	UMAP	23	0.89 ± 0.028	0.89 ± 0.010	0.89 ± 0.010

endosperm due to their high water content (Fig. 1C). Hereafter, we will refer to these embryo-endosperms simply as seeds. Thus, we digitally isolated all the seeds and obtained a collection of 3121 seeds in total. We used an in-house scipy-based python script to automate the image processing pipeline for all panicles and grains.

All the seeds were aligned with respect to their first three principal components. Since all the seeds are oblong in shape, this PCA-based alignment corresponds to the proximal-distal, medial-lateral, and adaxial-abaxial axes respectively (Fig. 1D). With this uniform alignment we were able to measure the length, width, heights, surface area and volume of each seed (Fig. 1E). In total, 11 different traditional shape descriptors were measured.

On the other hand, we computed topological shape descriptors using the ECT (Fig. 1F). For topological purposes, we treated each voxel-based image as a dual cubical complex where each nonzero voxel is treated as a vertex.[16] We used 158 different directions with 16 uniformly spaced thresholds (Fig. 1G). We emphasized directions toward the crease of the seeds. This yielded a 2528-dimensional vector for every seed. These high-dimensional vectors were later reduced in dimension using UMAP.[17]

The descriptiveness of both traditional and topological measurements was tested by training three non-linear support vector machines (SVM). These classified the seeds from the 28 distinct barley accessions based on three different collections of descriptors: traditional, topological, and combined. All the descriptors were centered and scaled to variance 1 prior to SVM. We randomly sampled 75% of the seeds from every accession as training data, and used the remaining 25% to test the accuracy of our classification model. This setup was repeated 100 times. Average scores were considered for the overall data set (Table 1), and for individual accessions (Fig. 2A).

3. RESULTS AND CONCLUSIONS

The majority of the barley accessions studied are more easily distinguished with the topological lens but not with traditional measures, with few exceptions (Fig. 2A). Exceptions like Hannchen, Han River and Palmella Blue have slightly distinctive traditional trait distributions, so it is important to take seed size into account. At the same time, we observe accessions like Alpha, Glabron, Minia, and Wisconsin Winter are poorly differentiated with traditional information but report considerably higher classification accuracies whenever using topological information. Classification accuracy is increased when combined with size-related information.

We performed Kruskal-Wallis one-way analyses[18] to determine if the Euler characteristic inter-accession variance was significantly different from the intra-accession variance at a particular slice and direction. The most accession-discerning slices and directions are close to the north and south poles (Fig. 2B), which correspond to the morphology of the crease and the bottom of the seed (Fig. 2C). That is, the topological shape descriptors do measure a morphological feature not explicitly measured by our traditional setting. These results follow a conservative 10^{-10} false discovery rate after considering a multiple test Benjamini-Hochberg correction.[19]

The Euler characteristic is a simple yet powerful way to reveal features not readily visible to the naked eye. There is "hidden" morphological information that traditional and geometric morphometric methods are missing. The Euler characteristic, and Topological Data Analysis in general, can be readily computed from any given image data, which makes it a versatile tool to use in a vast number of biology-related applications. TDA provides a comprehensive framework to detect and compare morphological nuances, nuances that traditional measures fail to capture and that remain unexplored using simple geometric methods. In the specific case of barley seeds presented here, these "hidden" shape nuances provides enough information to not only characterize specific accessions, but the individual spikes from which seeds are derived. Our results suggest a new exciting path, driven by morphological information alone, to explore further the phenotype-genotype relationship.

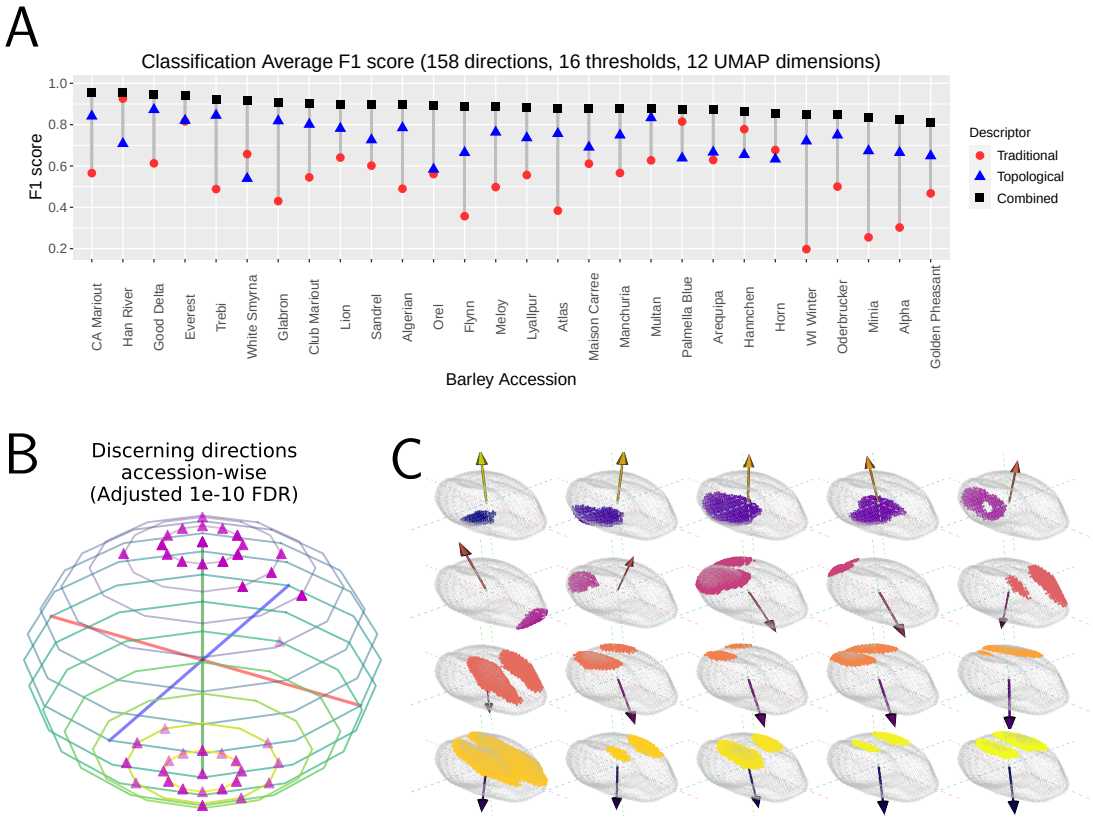

Figure 2. **SVM classification results for individual barley accessions and relevant ECT descriptors. (A)** Results when using a UMAP 12-dimension reduced topological vector. Accessions ordered according to their classification accuracy determined by the combined shape descriptors. **(B)** We examine the inter-accession and intra-accession variance differences of the Euler characteristic for each direction and threshold. Kruskal-Wallis analyses suggests that discerning directions and thresholds are mostly concentrated around the poles, and **(C)** correspond to the seed's crease and bottom morphology. Colors bear no special meaning.

SOFTWARE AND DATA AVAILABILITY

The processed and cleaned barley panicles and barley seeds X-ray CT 3D reconstructions can be found in the Dryad repository https://doi.org/10.5061/dryad.rxwdbrv93 (currently under curation process.) All of our code is available at https://github.com/amezqui3/demeter/. A collection of Jupyter notebook tutorials is also provided to ease the usage of the different components of the data processing and data analyzing pipelines.

ACKNOWLEDGMENTS

Dan Chitwood is supported by the USDA National Institute of Food and Agriculture, and by Michigan State University AgBioResearch. Elizabeth Munch is supported in part by the National Science Foundation through grants CCF-1907591 and CCF-2106578. Jacob Landis was supported by the NSF Plant Genome Postdoctoral Fellowship 1711807. Daniel Koenig is supported by an award from the National Science Foundation Plant Genome Research Program (IOS-2046256) and funding from the USDA NIFA (CA-R-BPS-5154-H).

REFERENCES

[1] Lestrel, P. E., ed., [*Fourier Descriptors and their Applications in Biology*], Cambridge University Press, Cambridge (1997).

[2] Kuhl, F. P. and Giardina, C. R., "Elliptic Fourier features of a closed contour," *Computer Graphics and Image Processing* **18**(3), 236 – 258 (1982).

[3] Lum, P. Y., Singh, G., Lehman, A., Ishkanov, T., Vejdemo-Johansson, M., Alagappan, M., Carlsson, J., and Carlsson, G., "Extracting insights from the shape of complex data using topology," *Scientific Reports* **3** (Feb. 2013).

[4] Munch, E., "A user's guide to topological data analysis," *Journal of Learning Analytics* **4**, 47–61 (07 2017).

[5] Amézquita, E. J., Quigley, M. Y., Ophelders, T., Munch, E., and Chitwood, D. H., "The shape of things to come: Topological data analysis and biology, from molecules to organisms," *Developmental Dynamics* **249**(7), 816–833 (2020).

[6] Turner, K., Mukherjee, S., and Boyer, D. M., "Persistent homology transform for modeling shapes and surfaces," *Information and Inference* **3**, 310–344 (12 2014).

[7] Richardson, E. and Werman, M., "Efficient classification using the Euler characteristic," *Pattern Recognition Letters* **49**, 99 – 106 (2014).

[8] Curry, J., Mukherjee, S., and Turner, K., "How many directions determine a shape and other sufficiency results for two topological transforms," (2018).

[9] Li, M., An, H., Angelovici, R., Bagaza, C., Batushansky, A., Clark, L., Coneva, V., Donoghue, M. J., Edwards, E., Fajardo, D., Fang, H., Frank, M. H., Gallaher, T., Gebken, S., Hill, T., Jansky, S., Kaur, B., Klahs, P. C., Klein, L. L., Kuraparthy, V., Londo, J., Migicovsky, Z., Miller, A., Mohn, R., Myles, S., Otoni, W. C., Pires, J. C., Rieffer, E., Schmerler, S., Spriggs, E., Topp, C. N., Van Deynze, A., Zhang, K., Zhu, L., Zink, B. M., and Chitwood, D. H., "Topological data analysis as a morphometric method: Using persistent homology to demarcate a leaf morphospace," *Frontiers in Plant Science* **9**, 553 (2018).

[10] Migicovsky, Z., Li, M., Chitwood, D. H., and Myles, S., "Morphometrics reveals complex and heritable apple leaf shapes," *Frontiers in Plant Science* **8**, 2185 (2018).

[11] Li, M., Frank, M. H., Coneva, V., Mio, W., Chitwood, D. H., and Topp, C. N., "The persistent homology mathematical framework provides enhanced genotype-to-phenotype associations for plant morphology," *Plant Physiology* **177**(4), 1382–1395 (2018).

[12] Diaz-Gárcia, L., Covarrubias-Pazaran, G., Schlautman, B., Grygleski, E., and Zalapa, J., "Image-based phenotyping for identification of QTL determining fruit shape and size in american cranberry (*Vaccinium macrocarpon L.*)," *PeerJ* **6**(e5461) (2018).

[13] Harlan, H. V. and Martini, M. L., "A composite hybrid mixture," *Agronomy Journal* **21**(4), 487–490 (1929).

[14] Harlan, H. V. and Martini, M. L., [*Problems and results in barley breeding*], US Department of Agriculture, Washington, DC (1936).

[15] Harlan, H. V. and Martini, M. L., "A study of methods in barley breeding," Tech. Rep. 720, US Department of Agriculture, Washington, DC (Feb. 1940).

[16] Wagner, H., Chen, C., and Vuçini, E., [*Efficient Computation of Persistent Homology for Cubical Data*], 91–106, Springer Berlin Heidelberg, Berlin, Heidelberg (2012).

[17] McInnes, L., Healy, J., and Melville, J., "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," (2020).

[18] Kruskal, W. H. and Wallis, W. A., "Use of ranks in one-criterion variance analysis," *Journal of the American Statistical Association* **47**(260), 583–621 (1952).

[19] Benjamini, Y. and Hochberg, Y., "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)* **57**(1), 289–300 (1995).