Characterizing the Role of Environment on Phenotypic Traits using Topological Data Analysis

Methun Kamruzzaman
School of EECS
Washington State University
Pullman, WA, USA
methun@eecs.wsu.edu

Ananth Kalyanaraman
School of EECS
Washington State University
Pullman, WA, USA
ananth@eecs.wsu.edu

Bala Krishnamoorthy
Department of Mathematics
Washington State University
Vancouver , WA, USA
bkrishna@math.wsu.edu

ABSTRACT

Phenomics is an emerging area within modern biology, which uses high throughput phenotyping tools to capture multiple environment and phenotypic trait measurements, at a massive scale. Due to the relatively nascency of the field, current tools and techniques used for phenomics data analysis are still, at large, the same tools originally designed to decode genotype to phenotype association studies (referred to as Genome Wide Association Studies). However, one of the key contributors to phenotypes (along with genotypes) is the environment. Yet there are presently no tools that allow users to analyze and characterize the role of environment in phenotypic performance. Here, we present a new algorithmic framework to characterize the role of environment on phenotypic traits. Our framework is an application of the Topological Data Analysis (TDA), which is an emerging branch in computational mathematics that deals with shapes and structures of complex data. To the best of our knowledge, this effort represents the first application of topological data analysis on phenomics data.

Categories and Subject Descriptors

I.3.5 [**Computational Geometry and Object Modeling**]; J.3 [**LIFE AND MEDICAL SCIENCES**]

General Terms

Algorithms, Design

Keywords

Computational phenomics, $G \times E = P$, topology data analysis, hypothesis extraction, visual analytics

1. INTRODUCTION

Understanding how *environments* impact various performance traits (*phenotypes*) of crop varieties (*genotypes*) is a core goal of modern biology. Different genotypes respond to different environments in distinct ways, making it challenging to select the most desirable varieties for a given environment. The phenotypic space is independently explosive in its dimensions. This emerging field of data gathering and subsequent analysis of phenotypic, genotypic and environmental data is collectively referred to as *phenomics* [3]. Figure 1 shows an example of a simple phenomics data set.

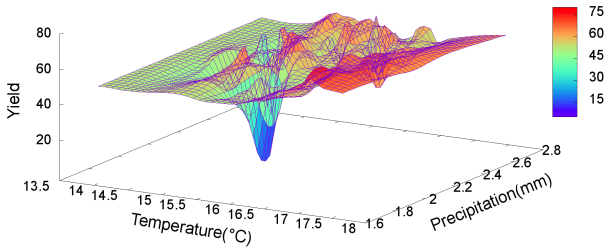


Figure 1: 3D representation of phenotype (yield) data with respect to environment (temperature and precipitation) for a real world data set [1].

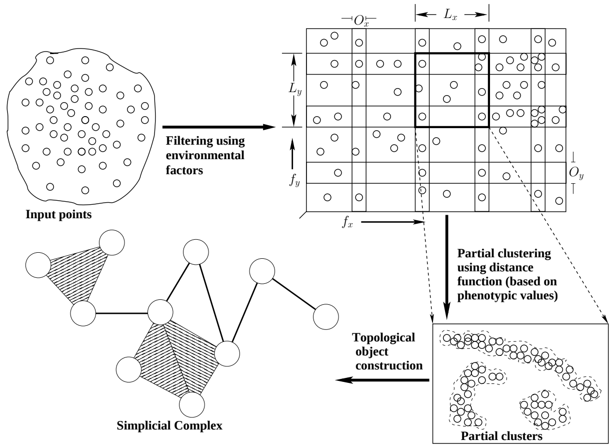
Contribution: We present an algorithmic framework using algebraic topology to study and characterize the effect of one or more environmental factors on a single phenotypic trait of interest. Our algorithmic framework is an implementation and application of the *Mapper* framework for conducting Topological Data Analysis (TDA), originally proposed by Gunnar Carlsson et al. [4]. A simplicial complex is a collection of simplices (nodes, edges, triangles, tetrahedra, etc.). In particular, each cluster is represented by a node (0-simplex). Whenever two clusters have non-empty intersection, we add an edge (1-simplex), and when three clusters intersect, we add a triangle (2-simplex), and so on.

2. METHODS

Let p be a *point* representing a crop individual that is measured at a particular time and location. Given an input set X of n points, the goal is to build a simplicial complex for X. A schematic illustration of the major steps of the algorithm is shown in Figure 2. These steps are:

Step 1: We define an open cover $\{U_i\}$ by introducing an interval length parameter L_i along each dimension. We also define an overlapping parameter O_i which is the overlap between two adjacent open cover $\{U_i\}$ and $\{U_{i+1}\}$. The interval lengths along all dimension creats a box. All the points p_i encompass in a box are stored in a hyper-octtrees.

BCB '16 October 02-05, 2016, Seattle, WA, USA
ACM 978-1-4503-4225-4/16/10.
http://dx.doi.org/10.1145/2975167.2985646.



O_x
L_x
L_y
f_y
O_y
f_x
Input points
Filtering using environmental factors
Partial clustering using distance function (based on phenotypic values)
Topological object construction
Partial clusters
Simplicial Complex

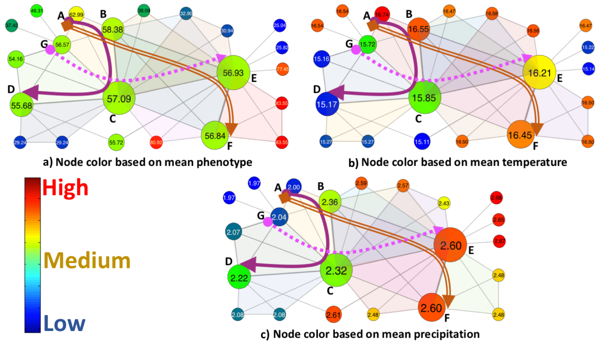
Figure 2: The TDA algorithmic framework for analyzing phenomic data.

Step 2: We apply density-based distance function DBSCAN [2] to cluster the phenotype x_i belong in p_i. Each cluster is treated as a node in a graph. All the clusters under the points those belongs in the overlapping area are shared and we draw an edge for shared area between two clusters. Hence, the clusters in the overlapping areas form a graph.

3. EXPERIMENTAL RESULTS

For testing our entire TDA framework, we used a real world data set [1] containing 7,454 (n) crop individuals, representing 180 soybean seed varieties, and cultivated across 350 sites from years 2008 through 2014 (shown in Figure 1).

One of our main results obtained by applying our TDA framework on the above data set is shown in Figure 3.



a) Node color based on mean phenotype

b) Node color based on mean temperature

c) Node color based on mean precipitation

Figure 3: Simplicial complex created by our TDA framework applied to the Syngenta data set. Nodes are labeled and colored by their mean phenotypic value in part (a), mean temperature value in part (b), and by mean precipitation value in part (c).

Path A-B-C-D: This path in Figure 3 identifies a collection of subpopulations (A,B,C,D) whose phenotypic performance is positively impacted by temperature, while precipitation does not have as much of an impact.

Path A-B-E-F: This path offers an example of subpopulations whose phenotypic performance is negatively impacted by precipitation (alone), while temperature does not have as much of an impact.

Path G-C-E: Along this path, the temperature values increase (part b) and the precipitation values also increase (part c) but the phenotypic values remain relatively unchanged which demonstrates the combined effect of both temperature and precipitation on the phenotypic values.

In addition, we also studied the effect of varying the length and overlap parameters on the filtering process of the TDA framework. We show below in Figure 4 the results for temperature. But the overall trends hold consistently for precipitation as well.

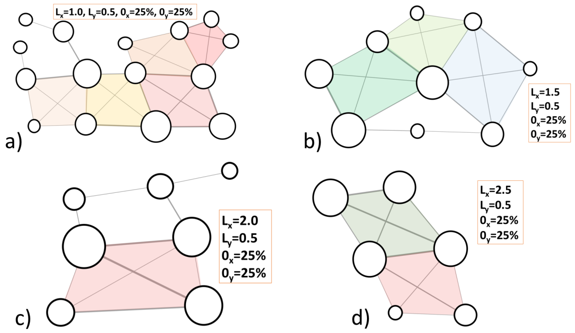


Figure 4: Effect of varying interval length (temperature) on resulting topological object.

4. CONCLUSIONS AND FUTURE WORK

We have presented a topological data analysis framework for phenomics. TDA is an emerging field within computational mathematics that deals with the study of complex shapes and structures in data. The TDA based approach presented in this paper aims to analyze the effect of environmental factors on phenotypic traits. Our approach inherits several key properties of TDA including a coordinate free approach, robustness to noise, and natural rendition to compressed representations of high-dimensional complex data spaces. Multiple future directions of research have been planned towards developing new algorithms for mining topological objects for hypothesis extraction.

5. REFERENCES

[1] Syngenta crop challenge in analytics. `https://www.ideaconnection.com/syngenta-crop-challenge/challenge.php`. [Last; accessed January-2016].

[2] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231, 1996.

[3] D. Houle, D. R. Govindaraju, and S. Omholt. Phenomics: the next challenge. *Nature Reviews Genetics*, 11(12):855–866, 2010.

[4] G. Singh, F. Memoli, and G. Carlsson. Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. In M. Botsch, R. Pajarola, B. Chen, and M. Zwicker, editors, *Proceedings of the Symposium on Point Based Graphics*, pages 91–100, Prague, Czech Republic, 2007. Eurographics Association.