bioRxiv preprint doi: https://doi.org/10.1101/159954; this version posted July 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Under Review

OXFORD

Data and Text Mining

Toward A Scalable Exploratory Framework for Complex High-Dimensional Phenomics Data

Methun Kamruzzaman [1,*], Ananth Kalyanaraman [2,*], Bala Krishnamoorthy [3] and Patrick Schnable [4,5]

[1]School of Electrical Engineering & Computer Science, Pullman, WA, 99164, USA, [2]School of Electrical Engineering & Computer Science, Pullman, WA, 99164, USA, [3]Department of Mathematics and Statistics, Vancouver, WA, 98686, USA, [4]Department of Agronomy, Ames, IA, 50011, USA, [5]Department of Genetics, Development and Cell Biology, Ames, IA, 50011, USA.

*To whom correspondence should be addressed.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Phenomics is an emerging branch of modern biology, which uses high throughput phenotyping tools to capture multiple environment and phenotypic trait measurements, at a massive scale. The resulting high dimensional data sets represent a treasure trove of information for providing an indepth understanding of how multiple factors interact and contribute to control the growth and behavior of different plant crop genotypes. However, computational tools that can parse through such high dimensional data sets and aid in extracting plausible hypothesis are currently lacking. In this paper, we present a new algorithmic approach to effectively decode and characterize the role of environment on phenotypic traits, from complex phenomic data. To the best of our knowledge, this effort represents the first application of topological data analysis on phenomics data.
Results: We applied this novel algorithmic approach on a real-world maize data set. Our results demonstrate the ability of our approach to delineate emergent behavior among subpopulations, as dictated by one or more environmental factors; notably, our approach shows how the environment plays a key role in determining the phenotypic behavior of one of the two genotypes.
Availability: Downloadable Source code and test data are freely available with instruction set at https://xperthut.github.io/HYPPO-X.
Contact: ananth@eecs.wsu.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

High-throughput technologies are beginning to change the way we observe and measure the natural world in life sciences. In medicine, physicians are using imaging and other specialized sampling devices to keep a longitudinal log of patients' drug/therapy response and other disease-related *phenotypes*. In agricultural biotechnology, phenotyping technologies such as cameras and LiDARs are being used to measure physiological and morphological features of a crop on the field. Further, advancements in *genotyping* technologies (sequencing) have made it possible to characterize and track genetic diversity and changes at a high resolution, and decode genetic markers key to performance traits. Taken together, advancements in these technologies are leading to a rapid explosion of high-dimensional data, obtained from a variety of sources.

A distinctive feature of these inherently high-dimensional data sets is that their generation is motivated more based on the availability and easy access to high-throughput technology (as opposed to specific working hypotheses). While there are some broad high-level questions or research themes that motivate the collection of data, the specific questions that relate to testable hypothesis and eventual discoveries (e.g., what genetic variations impact a physical trait, or how a combination of environmental variables control a phenotype) are *not* readily available *a priori*.

Consider the case of plant phenomics (6; 17). Understanding how different crop *genotypes* (G) interact with *environments* (E) to produce varying different performance traits (*phenotypes* (P)) is a fundamental goal of modern biology ($G \times E \rightarrow P$) (3; 21). To address this fundamental albeit broad goal, plant biologists and farmers have started to widely deploy an array of high-throughput sensing technologies that measure tens of crop phenotypic traits on the field (e.g., crop height, growth

© The Author 2015. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

bioRxiv preprint doi: https://doi.org/10.1101/159954; this version posted July 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

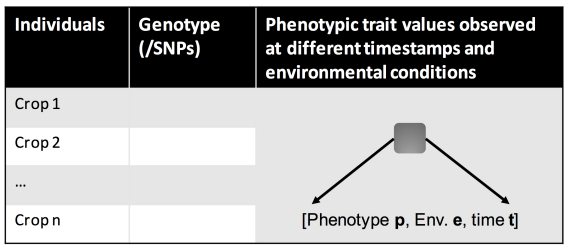

Fig. 1. Schematic table view of a multi-dimensional phenomics data set.

characteristics, photosynthetic activity). These technologies, comprising mostly of camera and other recording devices, generate a wealth of images (visual, infrared, thermal) and time-lapse videos, that represent a detailed set of observations of a crop population as it develops over the course of the growing season. Additionally, environmental sensors help in collecting daily field conditions that represent the growth conditions. Furthermore, through the use of sophisticated genotyping technologies, the genotypes of the different crop varieties are also cataloged. Fig. 1 shows a table view of a typical phenomics data set.

From this medley of plant genotypes, phenotypes, and environmental measurements, scientists aim to extract plausible hypotheses that can be field-tested and validated. However, the task remains significantly challenging, due to the dearth of automated software capabilities that are capable of handling complex, high-dimensional data sets. Scatter plots and correlation studies can reveal only high-level correlations and behavioral patterns/differences within data. However, it is common knowledge that different individuals or subgroups of individuals behave differently under similar stimuli. For instance, while it is useful to know that a given environmental variable (e.g., humidity) shows an overall positive correlation to a performance trait (phenotype), such high-level correlations obfuscate the variations within a population—e.g., how different subgroups or genotypes respond to different environmental intervals; or how one environmental variable interacts/interplays with another to affect the performance trait; or how the same genotype expresses variability in their performance in different environments. In fact, the need to identify such intra-population variations is what drives the generation of high-dimensional data in the area of phenomics (6; 17; 24).

1.1 Contributions

In this paper, we present a novel computational approach for extracting hypotheses from high-dimensional data sets such as phenomics collections. We formulate the problem of hypothesis extraction as one of: (a) identifying the key connected structural features of the given data, and (b) exploring the structual features in a way to facilitate extraction of plausible hypotheses.

Structure Identification: Our approach uses emerging principles from *algebraic topology* as the basis to observe and discern structural features from high-dimensional data. Algebraic topology is the field of mathematics dealing with the shape and connectivity of spaces (22; 8). There are multiple important properties of topology that make it particularly effective for extracting structural features from large, high-dimensional data sets (see Section 5).

Topological Object Exploration: While topological representations offer a compact way to represent and explore the data, the process of navigation and hypothesis extraction is an unexplored problem—one that is nontrivial with no current solutions. In this paper, we formulate this problem formally as one of identifying *maximal interesting paths* in a topological object. This novel formulation and a related algorithm allows our approach to systematically identify and evaluate potentially interesting features in topological representations.

Experimental Results: To demonstrate its effectiveness, we conducted a thorough experimental evaluation of our topological exploratory approach on a real-world maize data set, which contains environmental and phenotypic observations for two genotypes in two geographic locations (Kansas (KS) and Nebraska (NE)), over a period of 100 days—for a total of 400 "points". Our automated approach generates the topological object that clearly separates the genotypes of both locations. In fact, the interesting paths of the topological object identified by our method demonstrates the ability to: (i) identify the developmental and environmental stages of separation between the two genotypes, and between the two locations; and (ii) identify subtle variations in the behavior of individuals (or groups) within the subpopulations defined by genotypes and locations. Note that these findings are achieved in an entirely *unsupervised* manner by our approach.

We have implemented our approach as a software tool, which we call *Hyppo-X*[1]. The tool is available as open source on the following website: `https://xperthut.github.io/HYPPO-X`).

Even though we demonstrate the utility of our approach in the context of plant phenomics, our approach can be applied more broadly to other similar application contexts where the goal is to derive plausible hypotheses from complex, high-dimensional biological data sets.

The rest of the paper is organized as follows: Section 2 presents our algorithms and implementation details of our topologial exploratory framework. Section 3 presents the model to extract the hypothesis from the topological object constructed in Section 2 in the form of interesting paths. Section 4 presents a thorough experimental study and evaluation of our framework on a real-world maize data set. Section 5 reviews related tools used in phenomics research.

2 Building Compact Topological Representations

The first step in our approach is to construct topological representations using the connectivity properties of the data. The motivation is to obtain higher order structural information about the high-dimensional data prior to gleaning hypotheses. We adopt and adapt the Mapper algorithmic framework (25) for this purpose, and output our representations in the form of *simplicial complexes* (defined below). In what follows, we describe the details of our implementations of the individual steps of the framework. Fig. 2 is a schematic illustration of our approach.

Input: We are given a set of n points $\mathcal{S}$ in a d-dimensional space, representing the space of interest X. In the case of phenomics, a *point* $x \in \mathcal{S}$ represents a crop individual that is measured at a particular time t, and the dimensions represent the attributes which describe the point at that time. These include a set E of m factors (e.g., time, temperature, humidity, etc.), and a performance trait, the phenotype p (e.g., plant height or growth rate). Note that these dimensions represent continuous variables[2].

Output: We aim to create a highly compact coordinate-free representation of X as a *simplicial complex*, using a clustering (overlapping) of the points in X (represented by P here).

A *simplicial complex* is a collection of simplices (nodes, edges, triangles, tetrahedra, etc.) that fit together nicely—all subsimplices of each simplex are included in the collection, and any two simplices that intersect do so in a lower dimensional subsimplex. Specifically, each cluster is represented by a node (0-simplex). Whenever two clusters have a non-empty intersection, we add an edge (1-simplex), and when three clusters intersect, we add a triangle (2-simplex), and so on.

We now provide the main algorithmic details of the approach.

[1] Stands for "Hypothesis Extraction".

[2] A point may also have other non-continuous or static variables (e.g., the genotype). For the purpose of our topological representations we will use only the continuous variables.

bioRxiv preprint doi: https://doi.org/10.1101/159954; this version posted July 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

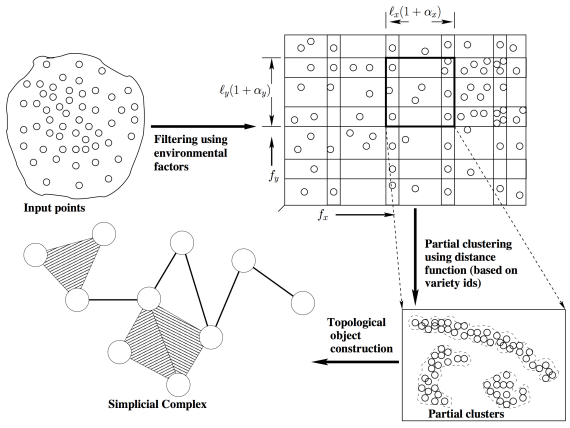

Fig. 2. The TDA algorithmic framework for analyzing phenomic data.

2.1 Filtering

The first component of the framework is a continuous function $f : X \to Z$ to a real-valued parameter space Z, called the *filter function*. For each factor Z_j, we define a filter function $f_j : X \to Z_j$. We generate the open cover $\mathcal{U}_j = \{U_{ij}\}$ of Z_j as follows:

(a)We divide each factor Z_j into n_j intervals ("sub-regions"), each of length ℓ_j. Thus the entire d-dimensional region is divided into $n_1 \times n_2 \times \cdots \times n_m$ subregions, where each subregion represents a hyper-rectangle of area $\ell_1 \times \ell_2 \times \cdots \times \ell_m$. Let the center of ith hyper-rectangle be $\{C_{1i}, C_{2i}, \ldots, C_{mi}\}$.

(b)We fix the center of each hyper-rectangle, and increase the length along each factor Z_j by a certain percentage α_j such that an overlapping region is created between consecutive pairs of the open sets U_{ij} and $U_{i+1,j}$, i.e., $U_{ij} \cap U_{i+1,j} \neq \emptyset$. After increasing the length of all sides in this fashion, the new area of the hyper-rectangle is $\ell_1(1+\alpha_1) \times \cdots \times \ell_m(1+\alpha_m)$[3]. A 2D example is shown in Fig. 2.

We formulate the efficient determination of individual point sets belonging to each hyper-rectangle as a problem of range querying. Specifically, we implement the following querying function:

Range Query: *Given X and a hyper-rectangle h, return the subset of points in X that lie in h.*

To run this query efficiently, we use k-dimensional hyper-octtrees (2; 10), which is a well known spatial data structure that uses recursive bisection to index a spatially distributed set of points. The compressed version of an n-leaves hyper-octree can be constructed in $O(n \log n)$ time (2). Once constructed, a balanced binary search tree that uses the order of the leaves is constructed. Using this auxiliary data structure, in combination with the hyper-octree, enables an $O(\log n)$ worst case search time for both point and cell searches (2). To answer the regional query for a hyper-rectangle h, we perform a top-down traversal of the hyper-octree by selectively retaining only those paths that can include at least one point within h. This can be achieved by keeping track of the corners of the cell defined by each internal node in the tree. This approach ensures that each such query can be answered in time that is bounded by the number of points in the hyper-rectangle.

2.2 Generation of Partial Clusters

Each open set (hyper-rectangle) computed by applying the filter functions is processed independently for generation of partial clusters. The goal of clustering is to partition the set of points in each hyper-rectangle based on their phenotypic performance.

Let U represent an open set of points $\{x_1, x_2, \ldots x_t\}$. Note that each point $x \in U$ has a phenotypic trait value denoted by $p(x)$. We define a *distance function* d based on the phenotypic values of points in U as follows. Given two points with trait values $p(x_i)$ and $p(x_j)$, the distance $d(i,j) = |p(x_i) - p(x_j)|$.

Given U and distance function d, a *partial clustering* is defined by a partitioning of the points in U. We denote the set of partial clusters resulting from any given open set U as $\mathcal{C}_U$. Subsequently, we denote the set of all partial clusters formed from *all* open sets (hyper-rectangles) by $\mathcal{C} = \bigcup_U \mathcal{C}_U$.

For the purpose of clustering, any distance-based clustering method can be applied. We implemented a density-based clustering approach very similar to that of DBSCAN (15). In the interest of space, we omit details of our implementation. Of note are, however, two key points: a) the set of partial clusters generated from within a hyper-rectangle represents a partitioning of those points; and b) two partial clusters generated from a pair of adjacent (overlapping) hyper-rectangles could potentially have a non-empty intersection in points. In fact it is this intersection that renders connectivity among the partial clusters generated, the information for which will be used in the subsequent step of simplicial complex generation.

2.3 Construction of Simplicial Complexes

From the set of partial clusters $\mathcal{C}$, we construct a simplicial complex K as follows. We describe the details for the 2D case, where no more than four open sets (hyper-rectangles) can mutually intersect. The extension to higher dimensions is straightforward. Starting with an empty simplicial complex, we implement the following steps:

i)A 0-simplex (or vertex) is added to the simplicial complex K for every partial cluster.

ii)Next, a 1-simplex (edge) is added to K for every non-empty 2-way intersection between any two partial clusters. Note that such intersections could exist only between partial clusters originating from different open sets.

iii)Following the same procedure as above, we also add 2-simplices (triangles) and 3-simplices (tetrahedra) to K by enumerating only those 3-way and 4-way intersections, respectively, that could be non-empty.

The required multi-way intersections are computed using the range querying function described earlier (in Section 2.1).

2.4 Persistent homology

We employ the concept of persistent homology (14) to choose the final topological object for further analysis. In particular, the method in which overlapping intervals are chosen (by specifying growing overlap percentages α_i, see Section 2.1) is already guided by this principle. Termed *multiscale mapper*, growing the intervals in this fashion ensures the topological objects formed (at each set of growing α_i values) satisfy a monotonic inclusion property (12). Hence results from persistent homology could be used to guarantee (theoretical) stability of the topological object formed (in the sense of persistence). At the same time, no implementation of multiscale mapper is known. Instead, we increase each α_i in steps of 2.5%, and construct the topological objects for each set of α_i values. We then construct the persistence barcodes (in dimensions 0, 1, and 2) using the sequence of topological objects formed by employing JavaPlex, a standard software tool for this purpose (1). We then pick α_i's such that all three barcodes do not change for values at or higher than the chosen cutoff, ensuring the corresponding topological object chosen is indeed stable. This chosen topological object is analyzed further to identify interesting paths.

[3] See Section 2.4 for further explanation of how we choose the α_i values.

bioRxiv preprint doi: https://doi.org/10.1101/159954; this version posted July 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

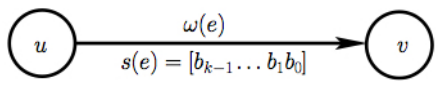

Fig. 3. An edge e between two intersecting partial clusters (nodes u and v). The direction of the edge indicates the direction in which the mean phenotypic/performance value increases. The signature $s(e)$ is a k-bit vector that captures the directions of change for each of the k filter functions (e.g., environmental variables) along the edge—0 implies reducing and 1 implies increasing. The ith bit corresponds to the ith filter function.

3 Interesting paths

The topological representation naturally reveals the underlying abstract structure of high-dimensional data. More specifically, a node (0-simplex) represents a partial cluster, which is a collection of points (i.e., a subpopulation) that shows similar phenotypic performance (by the distance function) under similar environmental conditions (filter intervals). An edge (1-simplex) connects two intersecting but distinct partial clusters. Therefore, by following a trail of nodes that show a monotonically varying performance, we can aim to capture the trail of subpopulations that gradually (or abruptly) alter their behavior under a continuously changing environment. Once identified, the user can extract points corresponding to these different subpopulations, and use them for comparative analyses and subsequent hypothesis extraction at the resolution of subpopulations.

Building on this idea, we formalize the notion of hypothesis extraction through exploration of topological objects as one of extracting "interesting paths" from the topological object. Details follow.

Using the simplicial complex K, we first construct a weighted directed acyclic graph $G(V, E)$, which represents the 1-skeleton of K along with some additional info. Let $V(K)$ and $E(K)$ denote the set of nodes and edges (0- and 1-simplices) of K. We set $V = V(K)$ and $E = E(K)$.

Note that each node $u \in V$ denotes a set of points that constitute a partial cluster in C. We denote this set as $X(u)$. Consequently, we assign real valued weights to all nodes and all edges in G—denoted by $\omega(u)$ and $\omega(e)$, respectively, for $u \in V$ and $e \in E$.

We set $\omega(u)$ to be the average phenotypic value for all points in u:

$$\omega(u) = \frac{\Sigma_{x \in X(u)} p(x)}{|X(u)|} .$$

For an edge $e = (u, v)$, we assign as its weight the absolute difference between the weights of the two nodes:

$$\omega(e) = |\omega(v) - \omega(u)| .$$

In addition, the direction of an edge e is set from the lower weight node to the higher weight node—i.e., if $\omega(u) \leq \omega(v)$, then $e : u \to v$; and $e : v \to u$ otherwise.

Edge and Path Definitions: If the simplicial complex was constructed using k out of the m continuous variables (as filter functions), then along each edge, each continuous variable Z_i can independently increase or decrease. Since we are trying to link the change of each of these variables relative to the change in phenotype (along an edge), we record a k-bit *signature* for each edge.

The *signature* of an edge $e \in E$ of $G(V, E)$, denoted by $s(e)$, is defined as a k-bit vector, where the ith bit is 1 if the direction of change for the continuous variable Z_i is consistent with the direction of the edge, and 0 otherwise. In other words, let an edge's direction be $u \to v$. Then, if the mean value for the continuous variable Z_i increases from u to v, then the corresponding signature bit is 1; and 0 otherwise.

Fig. 3 illustrates a directed, signed edge in our representation.

Let P denote a directed path in $G(V, E)$, containing a sequence of r edges $[e_1, e_2, \dots, e_r]$. Path P is said to be *exact* if the signature of all edges along the path is identical; and *inexact* otherwise.

We define the *interestingness score* $\mathcal{I}(P)$ of a path P as follows.

$$\mathcal{I}(P) = \Sigma_{i=1}^{r} \log(\text{rank}(e_i, P) + 1) \times \omega(e_i), \tag{1}$$

where the $\text{rank}(e, P)$ is the order of the edge e as it appears along the directed path P. Intuitively, we use the rank of an edge as an inflation factor for its weight—the later an edge appears in the path, the more its edge weight will count toward the interestingness of the path. This logic incentivizes the growth of long paths. The log function, on the other hand, helps contain this growth rate by treating edges that appear later in the path with comparable levels of importance (unless there is an order of magnitude increase in the path length).

We call two paths *overlapping* if they contain at least one edge in common.

Finding a set of Interesting Paths $(\mathcal{P})$ **problem:** Given $G(V, E)$ constructed as above, the goal is to *find a set of interesting paths* $\mathcal{P} = \{P | P \in \mathcal{P}\}$, such that (i) no two paths are overlapped and (ii) maximize $\mathcal{I}(\mathcal{P})$ =maximize $\sum_{P \in \mathcal{P}} \mathcal{I}(P)$.

3.1 The Algorithm

In this section, we present an efficient heuristic for the Interesting Paths problem. Our algorithm, while may not guarantee theoretical optimality in the interestingness score, runs in linear time (and space) and is effective in practice to identify meaningful paths with interesting aspects to them (as shown in Section 4). The main idea of our approach is to use dynamic programming.

Let e be an edge from $u \to v$. Then, we define the set of *candidate predecessor edges* for e as follows:

$$\text{Pred}(e) = \{f | f \in E, f : w \to u, \text{ such that } w \in V \text{ and } s(f) = s(e)\} \tag{2}$$

If the predecessor set is empty for an edge e, then the edge e is referred to as a *source edge*.

Let P_e denote an optimal path ending at edge e—i.e., a path with the maximum interestingness score ending at that edge. We compute two recurrences at edge e: i) the recurrence $T(e)$ as the interestingness score of P_e; and ii) a rank function $\text{rank}(e)$ that stores the rank of e along the path P_e. The recurrences are as follows.

$$T(e) = \begin{cases} \omega(e), \text{ if } e \text{ is a source edge,} \\ T(f^*(e)) + \log(\text{rank}(f^*(e)) + 1) \times \omega(e), \text{ otherwise,} \end{cases} \tag{3}$$

where, $f^*(e)$ is an *optimal predecessor* of e:

$$f^*(e) = \arg\max_f \{T(f) + \log(\text{rank}(f) + 1) \times \omega(e)\} . \tag{4}$$

Note that once $T(e)$ is computed, $\text{rank}(e)$ is given by:

$$\text{rank}(e) = \begin{cases} 2, \text{ if } e \text{ is a source edge,} \\ \text{rank}(f^*(e)) + 1, \text{ otherwise.} \end{cases} \tag{5}$$

A detailed pseudocode of our algorithm and a discussion on its runtime and memory complexities are provided in the Supplementary Document (Section S1).

4 Experimental Results

4.1 Experimental Setup

We tested our TDA framework on a real world maize data set. This data set consists of phenotypic and environmental measurements for two maize

bioRxiv preprint doi: https://doi.org/10.1101/159954; this version posted July 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

genotypes (abbreviated here for simplicity as A and B), grown in two geographic locations (Nebraska (NE) and Kansas (KS)). The data consists of *daily* measurements of the genotype's growth rate alongside multiple environmental variables, over the course of the entire growing season (100 days). For the purpose of our analysis we treat each "point" to refer to a unique [genotype, location, time] combination. Consequently, the above data set consists of 400 points (n). Here, "time" was measured in the Days After Planting (DAP).

Each point has one phenotypic value (growth rate) and 10 environmental variables (including but not limited to: humidity, temperature, rainfall, solar radiation, soil moisture, soil temperature). For the purpose of the studies presented in this paper, we select humidity, temperature and solar radiation, as they were the top variables that showed the most interesting trends and observations against the phenotypic trait. Results for other variables are omitted due to space considerations.

4.2 Topological object construction

We tested our framework using both single and two filter function(s).

Single filter function: The goal of our single filter study was to understand how the [location, genotype] combinations showed different performance in the phenotype as a function of time. Consequently, we used DAP as the filter function[4] and built our topological object, as described in Section 2.

Fig. 4 shows the resulting topological object. As can be seen, our method demonstrates the ability to clearly separate the two genotypes in both locations—as seen in the branching of paths, and the branch points in time indicate the DAP at which those genotypes separate. Notably, in both KS and NE, genotype B shows an accelerated growth rate earlier in its developmental stage than for A. Furthermore, both varieties in KS show an accelerated growth rate earlier than their NE counterparts. These observations are also confirmed by the scatter plot (Fig. S1 in the Supplementary Document).

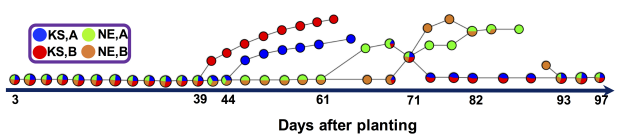

Fig. 4. Topological object based on the single filter function of days after planting (DAP). Here we considered both genotypes (A and B) in both locations (KS and NE). The pie chart at each node indicates the relative frequency of the four [location, genotype] combinations.

Two filter functions: While the above single filter function study shows some interesting behavioral differences between the genotypes with time and location, it is not adequate to provide any clues on the *basis* for such differences. In fact, what led to genotype B behaving differently across the two different locations? To delve into this specific question further, we include a second filter function, which can be any one of the environmental variables[5], and include only the genotype B points (i.e., [KS,B] and [NE,B]) for subsequent analysis. Fig. 5 shows the topological object generated using DAP and humidity as the two filter functions.

The topological object contains two large connected components. In the topological object colored by mean DAP (Fig. 5(C)), we notice that the largest connected component (right-top) captures the points from early to active growth stages, whereas the other connected component (left-bottom) holds the points for the post-growth stage (DAP roughly over 70).

4.3 Path analysis

As our next step, we identified multiple interesting paths from the topological object constructed using DAP and humidity as the two filter functions, using the method described in Section 3. Fig. 5 shows (using different colors) the paths that were automatically detected. For the purpose of this analysis, we kept the signature along a path fixed[6].

We now evaluate the qualitative significance of the interesting paths identified by our method, shown in Fig. 5:

➔The collection of co-located paths P_8, P_9, P_{10} essentially helps us understand how the genotype behaves in its early stages of development, in the two locations. More specifically, path P_8 starts with points from both locations because their performances in similar conditions (DAP and humidity) are also very similar; however, after roughly 20 DAP (Fig. 5C), the points from KS and NE separate (into P_{10} and P_9 respectively).

➔The sequence of paths $[P_6, P_5, P_1, P_3]$, which also includes the *most* interesting path by interestingness score (P_1), represents the active growth period for the KS population (see Fig. 5B). In this period, the growth rate increased from 1.38 cm/DAP to 9.73 cm/DAP, from approximately 35 DAP to 64 DAP (see Fig. 5C). In contrast, the plants in NE, despite being the same genotype, had very low growth rates during roughly the same period in time (35 DAP to 55 DAP; see Figs. 5B, 5C).

Incidentally, examining the humidity trends in the same period for these two locations (see Fig. 5D), we see that the humidity was very low in NE compared to KS, and that the increase in humidity values for the NE population (after 58 DAP) coincides with the increased activity in its growth rate (see Figs. 5C, 5D)—thereby giving us an indicator that humidity may have an active role in NE, perhaps more so than in KS, in accelerating growth rate during the mid-stages of development.

➔The sequence of paths $[P_7, P_2, P_4]$ represents the active growth period of the NE population, where the growth rate increases from 1.57 cm/DAP to 6.89 cm/DAP (Fig. 5B). This high activity period starts from approximately 60 DAP and ends roughly at 81 DAP. As indicated above, this active growth rate coincides with the period with higher humidity for NE.

To better understand the results and contrast it with more traditional approaches, we plotted all the genotype B points as a scatter plot, based on their DAP and humidity—see Fig. 6. The coloring of the points are by their location. As can be seen, the plot shows a clear separation between NE and KS humidity values, with NE exposed to lower humidity values than KS, in general. However, this is coarse-level information which can be easily obtained through a correlation test as well. At the same time, such tests are *not* adequate to provide meaningful insights into where the environmental or temporal triggers are relative to the performance, and how that behavior varies within a diverse population. That is where our topology-based approach can be useful—to make such inferences from the data and formulate testable hypothesis.

To better illustrate this advantage, we overlaid the interesting path sequences identified by our paths (discussed above) on to the scatter plot. These path sequences are shown as arcs in Fig. 6. As can be seen, our interesting paths show three major "features" within this scatter plot:

[4] The topological objects resulting from the use of other variables such as humidity and solar radiation as the single filter function are shown in the Supplementary Document (Section S2).

[5] Section S3.1 in the Supplementary Document shows the results for an alternative setup where the two filter functions used are two environmental variables.

[6] Our method has a feature to relax this condition, and we performed more analysis using that feature (not reported due to space). Of note is that the paths tend to get longer with this relaxation.

bioRxiv preprint doi: https://doi.org/10.1101/159954; this version posted July 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

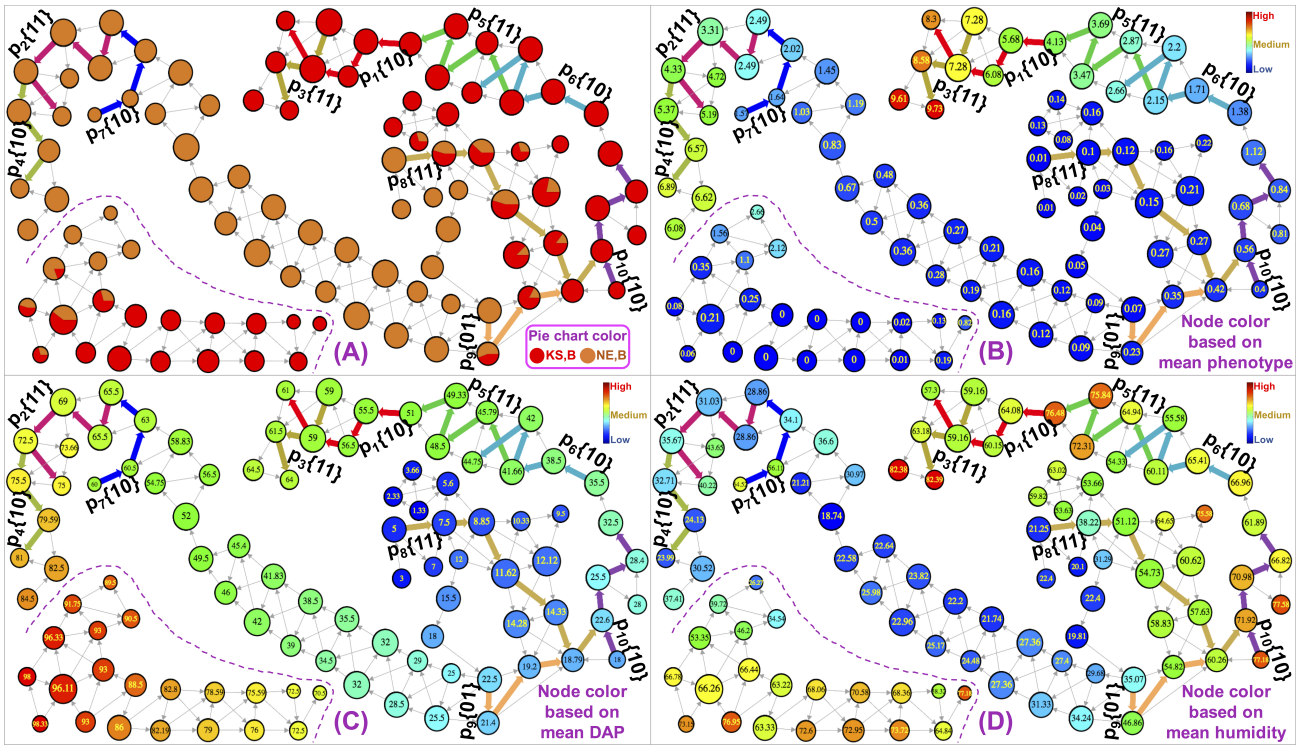

Fig. 5. The topological object constructed using DAP and humidity as the two filter functions, using only the [KS,B] and [NE,B] points. The same object is shown in the four panes, albeit with different coloring schemes. (A) shows each node (i.e., partial cluster) as a pie-chart of the relative concentrations of the two possible [location,genotype] combinations; (B) shows nodes colored by their mean phenotypic value; (C) shows nodes colored by their mean DAP value; and (D) shows nodes colored by their mean humidity value. The mean values are also indicated within the respective circles. The size of the circle for each node is proportional to the size of the corresponding partial cluster. Also shown highlighted as thick colored edges are the set of interesting paths identified by our method. Edge directions are from low to high mean phenotypic values. The interesting paths are labeled as $P_i\{s_1 s_2\}$, where i is the path number, and $s_1 s_2$ denotes the signature for that path (s_1 corresponds to DAP and s_2 corresponds to humidity). Recall that in the signature, 0 means decreasing and 1 means increasing.

i)the initial sequence where both NE and KS varieties behave similarly in their initial developmental stages, before branching out (around 20 DAP);

ii)the period of active growth for [KS,B] between roughly 35 and 65 DAP; and

iii)the period of active growth for [NE,B] appearing much later, between roughly 60 and 85 DAP.

More interestingly, the beginning of our interesting paths for [NE,B] is also for the first time the humidity value experienced a spike for that location—increasing from values under 30 to around 60s—effectively implying (or at least indicating) a probable cause for increased growth activity. After that trigger, minor fluctuations in humidity seemed to have little effect in the growth rate, which continued to increase through DAP 85. This study sets up a testable link between a genotype (B) and environmental variable (in this case, humidity) toward a performance trait (growth rate). Furthermore, the study raises a plausible working hypothesis that can be tested: "*If genotype [NE,B] is also exposed to a higher level humidity earlier on, during its developmental stage, then it is also likely to show an active growth rate earlier?*"

This illustrative example serves to demonstrate that our topology-based method also has the potential to enrich the information over what can be obtained through only conventional methods such as scatter plots. Note that our tool is meant for exploring high-dimensional data in a software-guided manner and more/other environmental variables can be included in our tests. In fact, in the Supplementary Document, we present more studies along this line by replacing humidity with temperature and solar radiation.

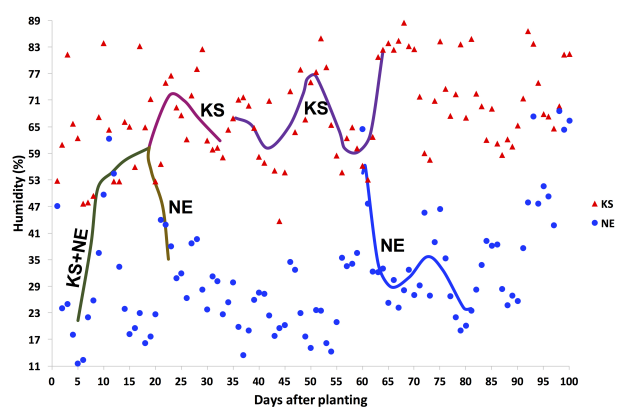

Fig. 6. Scatter plot of data points ([individual, date/time]) with respect to DAP and humidity in both locations. The color of a point indicates the corresponding location. According to this figure, the range of humidity is higher in KS compared to NE. The paths generated from our TDA framework are overlaid in this figure, which illustrate the phenotypic information.

Our study about the impact of temperature on plant growth rate is presented in supplementary document in Section S3.2.

5 Related work

We are not aware of any previous automated or semi-automated hypothesis extraction approaches for high-dimensional data sets.

bioRxiv preprint doi: https://doi.org/10.1101/159954; this version posted July 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Topology and Applications: There are several important properties that make algebraic topology particularly effective for gleaning structural features out of high-dimensional data. First, topology studies shapes in a *coordinate-free way*, which enables comparison among data sets from diverse sources or coordinate systems. Second, topological constructions are *not sensitive to small changes in data*, and robust against noise. Third, topology works with *compressed representations* of spaces in the form of *simplicial complexes* (or triangulations) (22), which preserve information relevant to how points are connected. Compared to more traditional techniques such as principal component analysis, multidimensional scaling, manifold learning, and cluster analysis, topological methods are known to be more sensitive to both large and small scale patterns (20).

Topological data analysis (TDA) has been applied to a wide range of application domains, albeit for mostly visualization purposes (11; 9; 4; 13; 18; 23; 26). The foundational work in TDA most relevant to this paper was done by Carlsson and coworkers (20). In (25), they describe a framework called *Mapper* to model and visualize high-dimensional data. Most of this work has been on the visualization front. A topology-based approach was also rated as the best overall entry at an expression QTL (eQTL) visualization competition organized by the BioVis community (5).

Tools for Plant Phenomics: Tools to decode associations between genotypes and phenotypes have been under development for over two decades. These tools look at the genetic variation observed at one or more loci along the genome and study their correlation to quantitative traits. The techniques used can be summarized as follows: i) Linkage mapping techniques that use prior knowledge on the location (markers) responsible for a certain trait; ii) Quantitative Trait Locus (QTL) mapping that extends linkage to an interval of co-located markers along the genome; and iii) Genome-Wide Association Studies (GWAS), which takes a whole genome approach by scoring multiple markers located across the genome for specific traits. In relation to capturing environmental variability, efforts have been sparse. Brown *et al.* (7) presented an experimental framework supplemented by GWAS to model environmental effects on phenotypes. Lou *et al.* (19) provided a generalized linear model-based method to capture gene to environment interactions.

6 Conclusion

We have presented a scalable exploratory framework for navigating high-dimensional data sets and applied it to plant phenomics data to analyze the effect of environmental factors on phenotypic traits. At its core, our approach is fundamentally different from state-of-the-art techniques in many ways: First, it inherits the advantages of topology including its use of a coordinate-free representation, robustness to noise, and natural rendition to compact representations. Second, by allowing the user to define multiple filter functions, it enables them to study the combined effect of multiple factors on target performance traits. Third, through its clustering and visualization capabilities, it provides a way for domain experts to readily observe emergent behavior among different groups or subpopulations without requiring the knowledge of any priors. This feature enables scientists to identify subpopulations, compare them, and perform more targeted studies to formulate and test hypotheses.

Our approach is scalable in that it can scale to large data sets containing possibly tens of thousands of points, reducing such large data to tens or hundreds of partial clusters, thereby making visualization and exploration possible. Although we have presented results on a smaller data set, we have tested our approach on larger data sets (e.g., with thousands of points (27); we did not present these results due to space constraints and also due to some missing information about the data (e.g., genotypes)).

While the scope of this work can be further expanded through application to a broader range of phenomics data collections, the results presented in this paper show a promising application of topology and its role in hypothesis extraction from high-dimensional data sets. Considering the nascency of the phenomics field, tools for users to explore data and help extract plausible hypothesis in a data-guided manner from large-scale complex data, will be important going forward.

Acknowledgements

We wish to thank Colton McNinch, who contributed to data collection. This research was in part supported by U.S. DOE grant DE-SC-0006516.

References

[1]Adams, H., Tausz, A., and Vejdemo-Johansson, M. (2014). javaplex: A research software package for persistent (co)homology. In *Mathematical Software–ICMS 2014*, volume 8592 of LNCS, pages 129–136. Springer. Software available at http://git.appliedtopology.org/javaplex/. [Last; accessed January-2017].

[2]Aluru, S. and Sevilgen, F. E. (1999). Dynamic compressed hyperoctrees with application to the n-body problem. In *Fou. Sft. Tch. The. Comp. Sci.*, pages 21–33.

[3]ASPB, R. (2013). Unleashing a decade of innovation in plant science: A vision for 2015-2025. Plant Science Research Summit.

[4]Ban, Y.-E. A., Edelsbrunner, H., and Rudolph, J. (2006). Interface surfaces for protein-protein complexes. *J. ACM*, **53**(3), 361–378.

[5]Bartlett, C. W., Cheong, S. Y., Hou, L., Paquette, J., Lum, P. Y., JÃ¤ger, G., Battke, F., Vehlow, C., Heinrich, J., and Nieselt, K. (2012). An eQTL biological data visualization challenge and approaches from the visualization community. *BMC bioinformatics*, **13**(Suppl 8), S8. 00005.

[6]Bilder, R. M., Sabb, F., Cannon, T., London, E., Jentsch, J., Parker, D. S., Poldrack, R., Evans, C., and Freimer, N. (2009). Phenomics: the systematic study of phenotypes on a genome-wide scale. *Neuroscience*, **164**(1), 30–42.

[7]Brown, T. B., Cheng, R., Sirault, X. R., Rungrat, T., Murray, K. D., Trtilek, M., Furbank, R. T., Badger, M., Pogson, B. J., and Borevitz, J. O. (2014). TraitCapture: genomic and environment modelling of plant phenomic data. *Current opinion in plant biology*, **18**, 73–79.

[8]Carlsson, G. (2009). Topology and data. *Bulletin of the AMS*, **46**(2), 255–308.

[9]Carlsson, G., Zomorodian, A., Collins, A., and Guibas, L. (2004). Persistence barcodes for shapes. In *ACM Sym. Geom. Proc.*, SGP '04, pages 124–135, ACM.

[10]Clarkson, K. L. (1983). Fast algorithms for the all nearest neighbors problem. In *24th Ann. Sym. Found. of Comp. Sci. (FOCS)*, pages 226–232. IEEE.

[11]de Silva, V. and Ghrist, R. (2007). Coverage in sensor networks via persistent homology. *Algebraic and Geometric Topology*, **7**, 339–358.

[12]Dey, T. K., Mémoli, F., and Wang, Y. (2016). Multiscale mapper: Topological summarization via codomain covers. In *27th Ann. ACM-SIAM Sym. Disc. Algo.*, SODA '16, pages 997–1013, SIAM.

[13]Edelsbrunner, H. and Koehl, P. (2005). The geometry of biomolecular solvation. In *Comb. Comp. Geom.*, in Volume 52 of *MSRI Publications*, pages 243–275.

[14]Edelsbrunner, H., Letscher, D., and Zomorodian, A. (2002). Topological persistence and simplification. *Disc. Comp. Geom.*, **28**, 511–533.

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

[16]Ghrist, R. (2008). Barcodes: The persistent topology of data. *Bull. AMS*, **45**, 61–75.

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

[18]Huan, J., Bandyopadhyay, D., Wang, W., Snoeyink, J., Prins, J., and Tropsha, A. (2005). Comparing graph representations of protein structure for mining family-specific residue-based packing motifs. *J. Comp. Biol.*, **12**(6), 657–671.

[19]Lou, X.-Y., Chen, G.-B., Yan, L., Ma, J. Z., Zhu, J., Elston, R. C., and Li, M. D. (2007). A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. *The American Journal of Human Genetics*, **80**(6), 1125–1137.

[20]Lum, P. Y., Singh, G., Lehman, A., Ishkanov, T., Vejdemo-Johansson, M., Alagappan, M., Carlsson, J. G., and Carlsson, G. (2013). Extracting insights from the shape of complex data using topology. *Scientific Reports*, **3**(1236).

[21]Martin, C. (2013). The plant science decadal vision. *Plant Cell Online*, **25**(12), 4773–4774.

[22]Munkres, J. R. (1984). *Elements of Algebraic Topology*. Addison–Wesley.

[23]Nicolau, M., Levine, A. J., and Carlsson, G. (2011). Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *PNAS*, **108**(17), 7265–7270.

[24]NIFA-NSF (2011). Phenomics: Genotype to Phenotype. NIFA - NSF Phenomics Workshop Report.

[25]Singh, G., Memoli, F., and Carlsson, G. (2007). Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. In *Sym. Point Based Graphics*, pages 91–100, Eurographics Association.

bioRxiv preprint doi: https://doi.org/10.1101/159954; this version posted July 6, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

[26]van de Weygaert, R., Vegter, G., Edelsbrunner, H., Jones, B. J., Pranav, P., Park, C., Hellwing, W. A., Eldering, B., Kruithof, N., Bos, E. P., Hidding, J., Feldbrugge, J., ten Have, E., van Engelen, M., Caroli, M., and Teillaud, M. (2011). Alpha, betti and the megaparsec universe: On the topology of the cosmic web. In *Trans. Comp. Sci. XIV*, volume 6970 of *LNCS*, pages 60–101. Springer.

[27]Syngenta Crop Challenge in Analytics, https://www.ideaconnection.com/syngenta-crop-challenge/challenge.php. [Last; accessed January-2017].