Topological Data Analysis in Population Genomics

Harry Emeric
Stanford University
harryem@stanford.edu

June 14, 2020

1 Introduction

A big challenge in Computational Biology is how to describe data in a high dimensional space, and much progress has been made in clustering, and dimensionality reduction techniques such as Principal Components Analysis, t-SNE, and UMAP [1]. These allow projections onto two dimensions which produce visualizations from which much of the structure of the space can be explained. However, a lot of information is lost in this process and its difficult to know what other interesting relationships are left unexplained. As such a more robust method of analysis would yield significant benefits.

Algebraic Topology is the study of invariants in spaces up to Homeomorphism, or smooth deformation, and Topological Data Analysis uses these concepts to interpret the shape of a dataset. By describing a space via an abstraction called a *Simplicial Complex*, that is a space constructed by gluing vertices together to make edges, gluing edges together to make triangles and so on, it is possible to compute algebraic invariants in a way well suited to algorithmic computation. From an evolutionary perspective, given the presence of phenomena such as lateral transfer of DNA between generations, the Tree structure where the parent(s) pass down DNA in a linear fashion does not appear to be a good model, and a graphs with loops would be better.

This paper seeks to analyze a genomic dataset with various software implementations of Topological tools, to see if these invariants appear, and see how the resulting relationships between samples (individuals) can help explain patterns and structure within and across populations.

2 Related Work

Recent work on analyzing genetic structure and how it relates to the underlying populations include finding deep substructure amongst indigenous Mexican populations [2], and the two dimensional projection of samples across 30 European countries is shown to correspond to their geographic origin [3]. Work in improving dimensionality reduction techniques for PCA [4] and UMAP [5] also address shortcomings in issues arising from current methods for analysis both within and between datasets.

3 Dataset

The data consists of genome samples taken from 2,748 individuals in the Himalayan region across 140 populations and 15 language families, and was collected to have good coverage of the region. They are unpublished samples collected by Dr Aashish Jha.

4 Methods

4.1 Preprocessing

The data is stored in .vcf format from the plink [6] package, which is converted to a genotype array, a boolean array of the expression of 119,875 alleles.

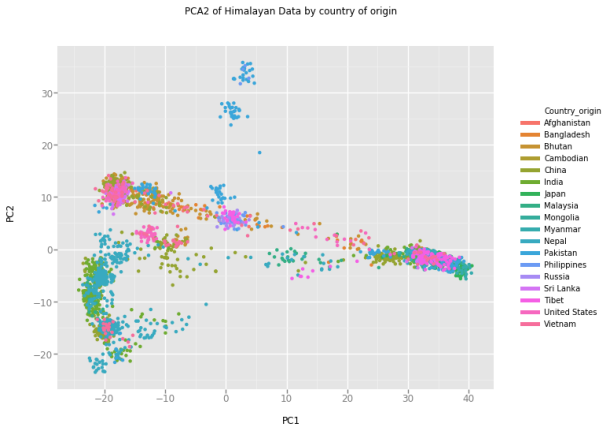


Figure 1: PCA2 plot with points labelled by country of origin.

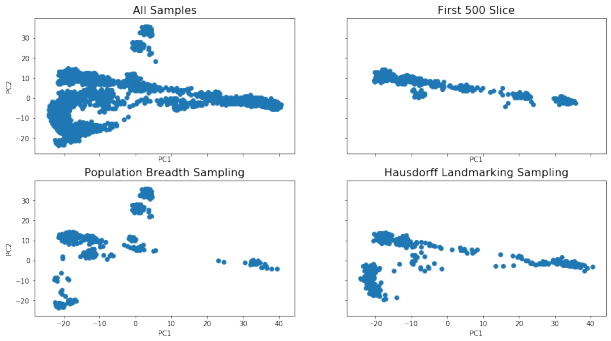


Figure 2: PCA2 plots comparing the full dataset with various methods of subsetting 500 points from the data. To demonstrate how much strucutre is lost using a primitive approach a simple first 500 rows method is included, as well as the Population Breadth based and Hausdorff Landmarking based approach.

The algorithms used in the analysis required significant computation, and the maximum number of samples our hardware could process was 500 - 1000. As such we explored two main methods for subsetting the data. The first intended to maintain breadth across populations as much as possible. This worked by randomly adding one sample for all populations, until there were more populations than samples left to add to the subset, at which point a random set of populations was chosen first to complete the desired subset size.

The second method, Hausdorff landmarking, looked to maintain maximal distance of the subset. It does this by starting with one point, and proceeding in a stagewise greedy manner to add the point which reduces the Hausdorff distance between the selected and remaining points until there are the predetermined number of points.

4.2 Topological Data Analysis

Topological Data Analysis is a rapidly evolving field within Data Science and presents a more complete way to look at structure in high dimensions. This paper [7] is a more detailed introduction to the field and this book [8] describes current applications into genomics and evolution. Here we introduce certain key definitions.

Simplicial Complexes

A simplicial complex can be thought of as a generalization of a graph to a countable number of dimensions. A 0-simplex is a point or vertex, a 1-simplex is a line segment or edge connecting two points, a 2-simplex is a triangle made up of three connected edges, and so on. Simplicial complexes are useful to work with because we can express the boundaries of a simplex as a union of simplices and as such abstractly represent an object as a joined set of lower dimensional objects under a very restricted set of mapping functions.

Vietoris-Rips Complex

The dataset is a point cloud X in R^n, that is a finite topology. The Vietoris-Rips Complex, or Rips Complex, for a parameter ϵ is the simplicial complex with vertices in X and $\{x_0, x_1, ..., x_k\}$ spans a k-simplex if and only if $d(x_i, x_j) \leq \epsilon$ for all $0 \leq i, j \leq k$

Persistant Homology

A filtration is constructed by gradually increasing ϵ and tracking which k-simplic exist for that ϵ and track how long a simplex lasts. Initially points are connected via an edge as ϵ gets big enough, a loop may eventually form, and then be destroyed as the loop is "filled in". Persistant Homology studies how

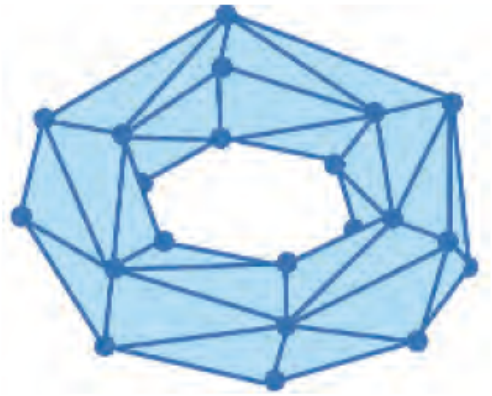


Figure 3: A torus triangulated as a simplicial complex. Source: Topological Data Analysis for Genomics and Evolution [8].

robust a k-simplex is to changes in ϵ, with a longer time from creation to destruction (called birth-death time) associated with structure that best describes the object.

Figure 4: An "ideal" rips filtration illustrates the concept of Persistant Homology. Initially in (A) ϵ is too low for the points to be connected, they then become connected via a loop, which eventually dies when it gets filled in. $\epsilon = \alpha$ is the birth and $\epsilon = \beta$ is the death of the loop, and the birth-death bardcode is shown in (B). Source: Topological Data Analysis for Genomics and Evolution [8].

4.3 Software Implementations

Software packages for this type of analysis fit into two categories, either very performant but lacking functionality, or better functionally but difficult to use on a large dataset such as ours. In the perfomant category I used the Ripser package [9] in python which was useful for exploratory purposes to get birth-death plots for various distance metrics and reducing dimensionality by taking a subset of the principal components. For improved functionality, using the results from Ripser to get the maximum threshold parameter, I used BATS.py [10] to plot the H_1 homology group on a two dimensional representation of the data obtained from the first two principal components.

5 Results

The Birth-Death diagrams produced by Ripser provide a plot of the Homology groups in the Rips filtration, each with persistence measured by the time between the birth and death of each simplicial complex. Due to computational restrictions we could only calculate to H_1. If a complex dies soon after its created it is not a robust description of the structure of the data, and will be plotted near the diagonal, that is the line birth = death. Presence of one or more points far above the diagonal give evidence that the Rips filtration identifies interesting and robust topological structure.

The full dataset was too big to run, so we looked at the Birth-Death plots for 2, 50, and all 5,496 principal components, and precalculating distance matrices for Euclidean, Manhattan, and Hamming distance using the full dataset. The matrix was normalized before calculating the principal component so the rows (gene expressions) summed to 0.

Seeing as the genoytype data is a boolean matrix with each expression with zero unless there is a mutation at that site, in which case it is one, the Manhattan distance measure seemed the natural choice between samples. The plot shows there is both one very persistent complex and a cluster of other complexes above the diagonal, which suggests the analysis has discovered the structure. Other distance measures did not yield significantly different results, each of which were at least slightly less compelling than the Manhattan distance. With a pre-computed distance matrix, Ripser took between 2-3 minutes to complete.

Given how widespread a two dimensional projection is in analyzing genomic data we thought it would

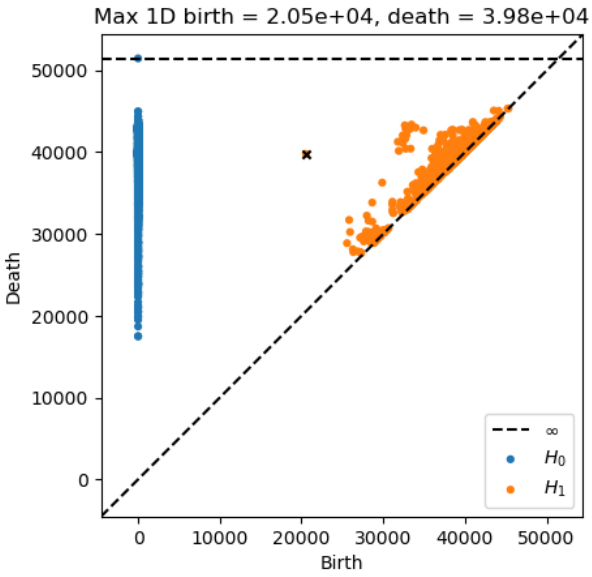
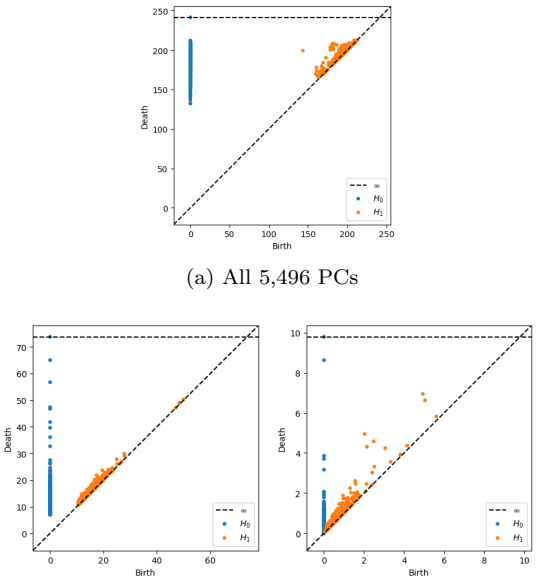


Figure 5: Ripser Birth-Death plot for H_0 and H_1 using Manhattan distance on all samples and using all dimensions.



(a) All 5,496 PCs

(b) 50 PCs

(c) 2 PCs

Figure 6: Ripser Birth-Death diagrams for various resolutions of PCs

be interesting to see how sequentially reducing the number of principal components would affect the Rips filtration. Rotating the original matrix to a square matrix unsurprisingly did not make much difference, but in reducing to 50 PCs most of the structure is lost, as can be seen by most points being near the diagonal. Most interesting however is that in two dimensions there are many persistent structures which were not there in 50; the fact that structure disappears and then comes back is strong evidence that the structure seen in the two dimensional projection is not the same as the global structure.

As the distance matrices could be precomputed before being input into the analytics software, this meant that computational benefits from reducing the resolution of the data were limited, and the bottleneck was on the number of samples. BATS.py could process up to 900 samples before it reached its memory limits, which ran in 18 minutes for Hausdorff landmark subsetting process described in 4.1. Even for this much reduced dataset, a ring structure can be seen when the representative H_1 group is plotted against the 2D projection.

These representative plots also gave a much stronger suggestion that the Hausdorff subsetting method was superior to the population breadth method. A small ring can be seen for the latter in a subset of 700 samples, and is far less clear than that same size for Hausdorff.

6 Conclusion and Future Work

Our results provide evidence that patterns in genomic data can be discovered by the more robust set of tools given by Topological Data Analysis, and furthermore these appear to give better description of the geometric object that is a point cloud than a two dimensional projection.

We would look to expand on this paper by analyzing which individual samples are involved in each of the longest birth-death points, and incorporating the ethnographic information to make inferences about

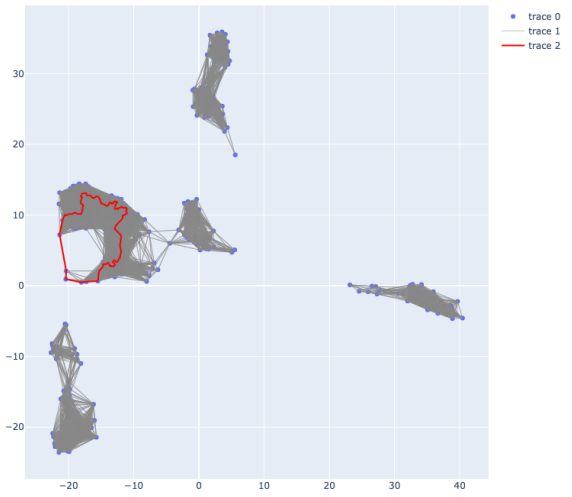
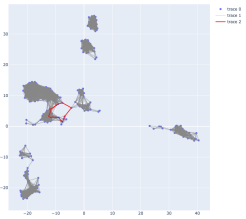


Figure 7: Representative Cocyles plot in 2 PCs for 900 samples obtained from Hausdorff Landmarking subsetting.

relationships between and within populations.

7 Acknowledgements

Thanks to Alex Ioannidis for mentoring the project, setting the vision and connecting the dots between Genomics and Algebraic Topology, to Brad Nelson for sharing his deep knowledge of the field of Topological Data Analysis as well as implementing certain new functionality and documentation in BATS.py, and to the Bustamante lab for providing the infrastructure and data.



(a) Subset of 700, Hausdorff Landmarking

(b) Subset of 700, Population Breadth Sampling

Figure 8: H_1 representative plots on 2 PCs of data

References

[1] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2018.

[2] Andrés Moreno-Estrada, Christopher R. Gignoux, Juan Carlos Fernández-López, Fouad Zakharia, Martin Sikora, Alejandra V. Contreras, Victor Acuña-Alonzo, Karla Sandoval, Celeste Eng, Sandra Romero-Hidalgo, Patricia Ortiz-Tello, Victoria Robles, Eimear E. Kenny, Ismael Nuño-Arana, Rodrigo Barquera-Lozano, Gastón Macín-Pérez, Julio Granados-Arriola, Scott Huntsman, Joshua M. Galanter, Marc Via, Jean G. Ford, Rocío Chapela, William Rodriguez-Cintron, Jose R. Rodríguez-Santana, Isabelle Romieu, Juan José Sienra-Monge, Blanca del Rio Navarro, Stephanie J. London, Andrés Ruiz-Linares, Rodrigo Garcia-Herrera, Karol Estrada, Alfredo Hidalgo-Miranda, Gerardo Jimenez-Sanchez, Alessandra Carnevale, Xavier Soberón, Samuel Canizales-Quinteros, Héctor Rangel-Villalobos, Irma Silva-Zolezzi, Esteban Gonzalez Burchard, and Carlos D. Bustamante. The genetics of mexico recapitulates native american substructure and affects biomedical traits. *Science*, 344(6189):1280–1285, 2014.

[3] John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R. Boyko, Adam Auton, Amit Indap, Karen S. King, Sven Bergmann, Matthew R. Nelson, Matthew Stephens, and Carlos D. Bustamante. Genes mirror geography within europe. *Nature*, 456(7218):98–101, 2008.

[4] Abubakar Abid, Martin J. Zhang, Vivek K. Bagaria, and James Zou. Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nature Communications*, 9(1):2134, 2018.

[5] Ben-Eghan C Gravel S. Diaz-Papkovich A, Anderson-Trocmé L. Umap reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts. *PLoS Genet.*, 2019.

[6] Plink is a free, commonly used, open-source whole-genome association analysis toolset designed by shaun purcell.

[7] Gunnar Carlsson. Topology and data. Technical report, 2008.

[8] Raul Rabadan and Andrew J. Blumberg. *Topological Data Analysis for Genomics and Evolution: Topology in Biology*. Cambridge University Press, 2019.

[9] Christopher Tralie, Nathaniel Saul, and Rann Bar-On. Ripser.py: A lean persistent homology library for python. *The Journal of Open Source Software*, 3(29):925, Sep 2018.

[10] Brad Nelson. Basic applied topology subprograms.