Project Report: Detection of Topological Features on Gene Expression Data

Minjian Li
School of Electrical Engineering and Computer Science, Washington State University
minjian.li@wsu.edu

***Abstract*—This report shows the results of application of TDA on gene expression data. TDA includes a number of methods; they are mapper, persistence, triangulation, and so forth. Via applying mapper and persistence analysis, certain interesting correlative relationships between topological features and gene expression can be discovered.**

Keywords—Topological Data Analysis, Topological Features, Feature Detection, Gene Expression, HIV, H1N1, Helicobactor, Mapper, Persistent diagram

I. Introduction

TDA is an abbreviation standing for Topological Data Analysis. This analytical technique becomes more and more popular in the past few decades. Traditional data analytical methods in Data Science and Statistics, such as classification, clustering, and regressions, as well as statistical values, are weak to visualize data lying in high dimensions, while PDA is able to capture and help to unravel topological features, "holes" for instance, of high dimensional data. With extreme structure complexity, chromosomes usually contain hundreds of thousands of packages of genes. As a result, gene expression data are typically existing in a high-dimensional space. TDA, therefore, is constructive in genomic study, and it is used in this project. This project is an experimental project aiming at discovering whether hidden structures of gene expression data are correlated to the properties of diseases, by repeating the first part of TDA techniques mentioned in Svetlana Lockwood and Bala Krishnamoorthy's paper [1].

II. Data

Datasets (*GDS5294*, *GDS5411*, and *GDS4240*) used in this project can be found on "Gene Expression Ominibus" [2]. It is worth noting that gene expression data on the same topic with different values can probably lead to the same TDA result since several genes are able to express in the same way within the same topic and, therefore, form a same landscape in topological space. In order to avoid redundant data that cause meaningless results, GDS5294, GDS5411, and GDS4240 are three different data sets with topics in HIV-1 Vpr Protein, Helicobacter pylori(HB), and Peripheral blood cells under H1N1 respectively. As such three topics are well-known in bio-fields, I used them as target data in this project; they are also the top three biological topics that I am interested in.

Raw data are usually noisy. It could be untidy in many ways, such as missing values, useless columns, and one column with multiple variables, which disrupts the accuracy of analysis. As a result, only extract the columns that related to the project is of vital importance. Fortunately, such three datasets are already tidy by only finishing the extraction.

Considering gene expression data are generally lying in high dimensions, for example, *GDS5294* is a dataset belongs to R^{54675}, the computational complexity is enormously high. In order to avoid unnecessary calculations, all datasets have to be reshaped by dualization so that it is reduced to a lower dimensions. The universal expression of reshaped dataset is shown in (1),

$$v_i = \{x_1, x_2, \dots, x_j\} \in R^j \quad (1)$$

where i is the number of genes; v_i is the i-th genes; j is the number of samples, and x_j is the value of the j-th sample.

III. Methodology

Persistence and Mappers are main analytical methods applied on this project.

A. Mapper

Mapper analysis is one of the informative visualization methods on TDA. It captures the skeleton (nerve) of an object via separating the interval that covers the entire object into several connected or overlapping subintervals. Through the visualization of dataset, certain topological features can be intuitively discovered. In this project, I use Kepler Mapper to visualize all three datasets [3].

B. Persistent Diagram

With the purpose of persistent analysis, it is necessary to construct Vietoris-Rips complex for each dataset with pairwise Euclidian distance less than 2ξ, where ξ is an optimal non-zero constant. As ξ increases, the number of edges in Vietoris-Rips complex also increases. If the increment of edges is close to zero in an interval of ξ, then all ξ in such interval is optimal. This is the way I determine optimal ξ. On the basis of optimal Vietoris-Rips, a persistent diagram can be generated by Ripser in Python [4]. Due to the weak computational power of the device, for each dataset, only 12000 random samples are selected to construct Vietoris-Rips. If the number of samples is larger than 12000, the distance matrix will be out of memory.

IV. Results and Discussions

Since the limitation of the device, all results are observed by running multiple subsets, with size 12000, of each gene expression datasets.

A. GDS5294 ─ HIV-1 Vpr Protein

Figure 1 illustrates the approximate structure of dataset *GDS5294*. It forms a shape of net and its structure is partially compact with several loops locating on the "outskirt" of the net. Generally, large area of structure is compact and formed as a tetrahedral net, which is stable and difficult to be destroyed. Such stable structure probably implies that those part of genes do not be affected by HIV-1 Vpr protein. Several holes in the mapper probably mean HIV-1 Vpr protein is able to affect few genes expression.

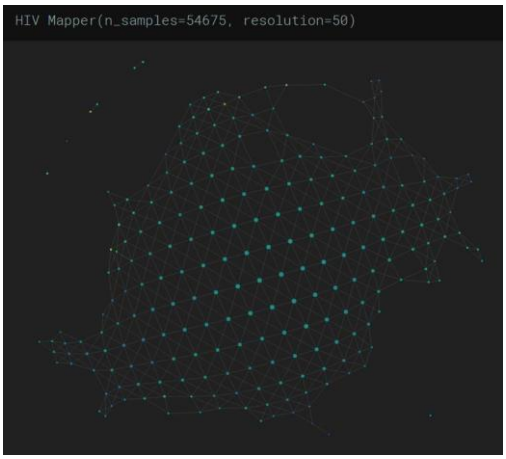

Fig.1 Mapper for GDS5294

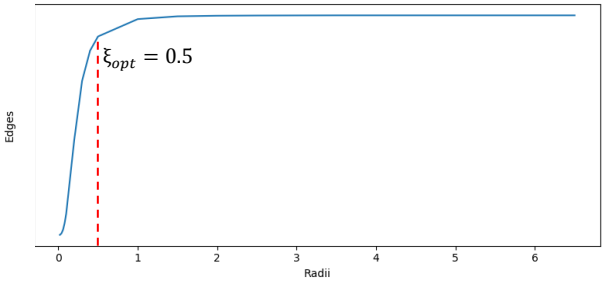

Fig.2 Relationship between ξ and edges of Vietoris-Rips (GDS5294)

By running multiple iterations with different value of ξ (see Figure 2), $\xi_{opt} \in [0.5, \infty)$ is the optimal value of ξ that generates a suitable Vietoris-Rips complex. As a result, let $\xi_{opt} = 0.5$ to get the persistent diagram (shown in Figure 3). Figure 3 is a representative of all persistent diagram generated in this section. Although multiple subsets are used for the generation, there are just slightly differences among the persistent diagrams. This perhaps infers the same hypothesis as what we get from mapper.

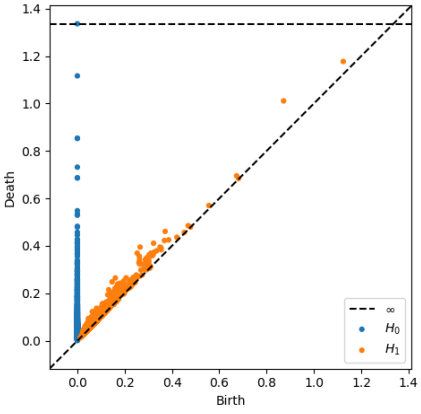

Fig.3 Persistent Diagram for H_0 and H_1 on GDS5294

B. GDS5411 ─ Helicobacter Pylori(HB)

Figure 4 illustrates the approximate structure of dataset *GDS5294*. It forms a shape of strip and its structure is incompact with a brunch of hollows and spikes, just like a sponge but long, flat and spiky. Notice that structure of the data set containing either non-infected gaster or infected gaster is smoother than the structure containing both infected and non-infected samples. This probably implies that the chaos of the structure could be an indicator for infection of Helicobacter.

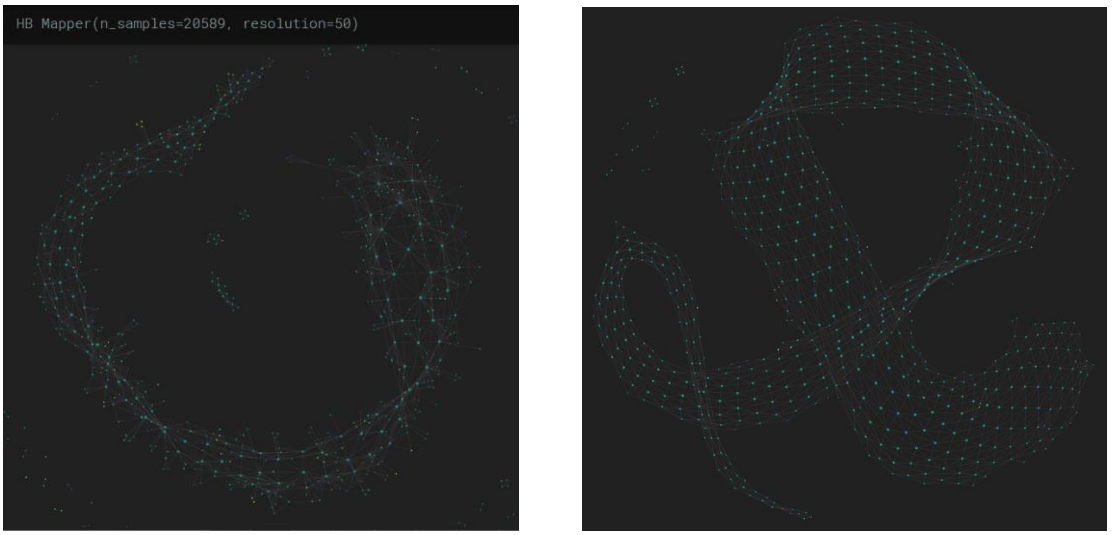

Fig.4 Mapper for all samples (left) and Mapper only for non-infected samples (right)

Similarly, according to Figure 5, the optimal of ξ for this dataset is $\xi_{opt} \in [7.5, \infty)$. Moreover, Vietoris-Rips complex for Helicobacter is not as sensitive as HIV. Let ξ_{opt} = 0.5, then we will have persistent diagram. Surprisingly, nothing outstanding is observed by comparing persistent diagram of infected and non-infected samples. However, if two kinds of datasets are collected together to generate a persistent diagram (see Figure 6), a dense area and a sparse area emerge in H_0 and H_1, which is similar to the situation in mapper.

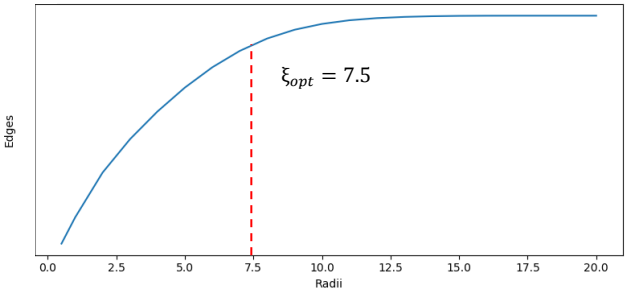

Fig.5 Relationship between ξ and edges of Vietoris-Rips (GDS5411)

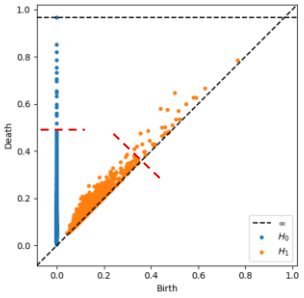

Fig.6 Persistent Diagram for H_0 and H_1 on GDS5411

C. GDS4240 ─ Peripheral Blood Cells under H1N1

The structure of *GDS5294* has the same general shape as Helicobacterthe; however, they can be distinguished in details. In Figure 7 is a mapper for gene expression of peripheral blood cells on healthy people. Intuitively, there are several "sprouts" on the strip. If a person becomes infected, there are more sprouts growing up on the strip, especially at the end of the strip (see Figure 8 and 9). In addition, as the duration of infection increases, the number of sprouts increases as well. Such observations probably suggest that the infection of H1N1 is positively correlated with the gene expression of peripheral blood cells.

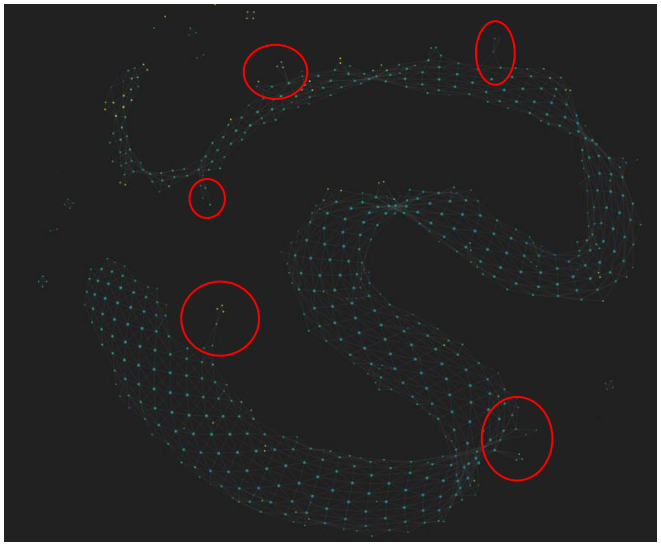

Fig.7 Mapper for GDS4240 (controllor)

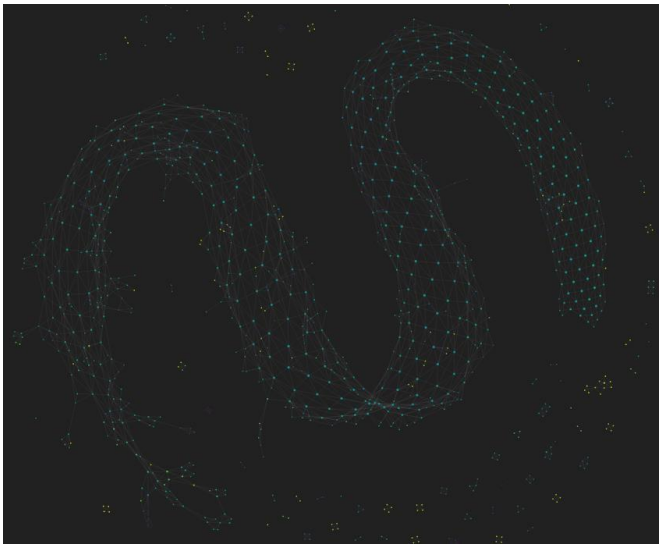

Fig.8 Mapper for GDS4240 (infected less than one day)

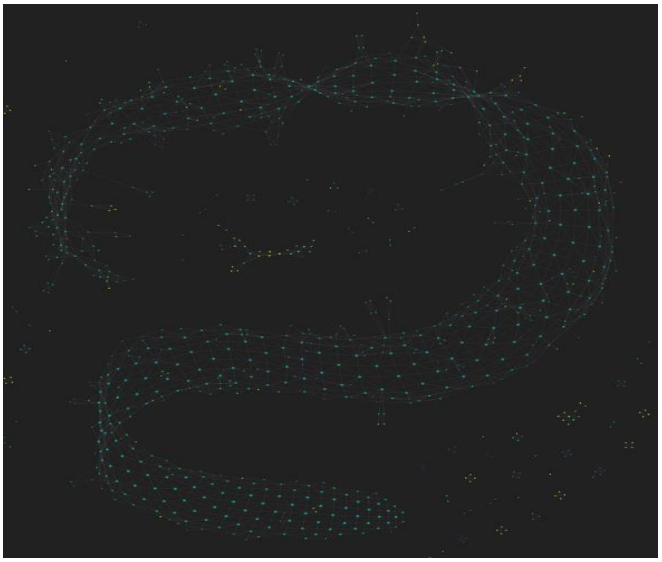

Fig.9 Mapper for GDS4240 (infected 6 days)

By repeat the same methods, the optimal of ξ for this dataset is $\xi_{opt} \in [12.5, \infty)$, which is shown in Figure 10. Moreover, Figure 11 illustrates that, for gene expression data of healthy people, nearly all features in H_1 are formed in a shorter interval of filtrations, while that for

infected people are formed in a longer interval. Also, several last-formed features in H_1, with regard to data of infected people, tend to have a longer life expectancy. Such differentiations indicate that the infection of H1N1 is highly correlate with gene expression of peripheral blood cells.

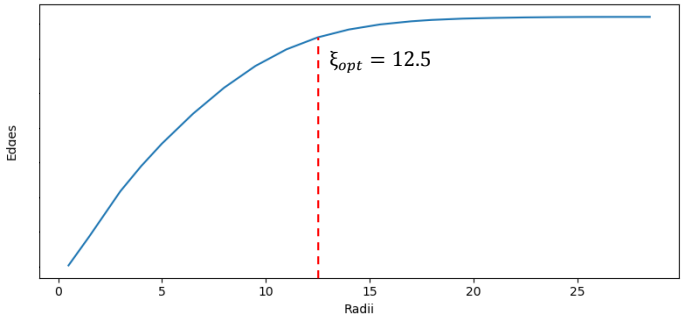

Fig.10 Relationship between ξ and edges of Vietoris-Rips (GDS4240)

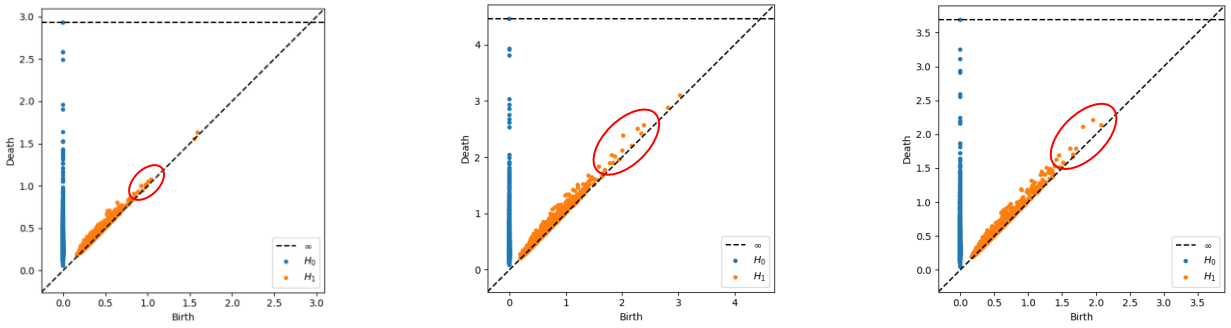

Fig.11 Persistent Diagram for H_0 and H_1 on GDS5411 which representing gene expression of people who is healthy (left), infected within 1 days (middle), and infected for 6 days (right)

V. Conclusion

According to the results and discussion section in this report, topological features could be a reflection of gene expression to a certain extend. For the selected datasets, the link between topological structure and gene expression can be successfully found via TDA. However, the limitation of computational ability of the device can probably cause a huge discrepancy of results. Moreover, the results could be different if another method in TDA can be applied on this project. Therefore, this project can be improved in such aspects. Furthermore, TDA can be not only used for gene expression data, it can be also used for other fields, such as computer science and physics. That is because we need to understand objects with high dimensionality as we learn deeper in our field, and TDA is the bridge connecting between high-dimensional space and our world space.

References

[1] S. Lockwood and B. Krishnamoorthy. “Topological features in cancer gene expression data”. In Paci_c Symposium on Biocomputing Co-Chairs, pp. 108-119. World Scienti_c, 2014.
[2] National Center for Biotechnology Information. Gene expression omnibus. Online; accessed 04-May-2018; https://www.ncbi.nlm.nih.gov/sites/GDSbrowser.
[3] Hendrik Jacob van Veen, and Nathaniel Saul. (2017, November 17). MLWave/kepler-mapper: 186f (Version 1.0.1). Zenodo. http://doi.org/10.5281/zenodo.1054444
[4] Tralie et al., (2018). Ripser.py: A Lean Persistent Homology Library for Python. Journal of Open Source Software, 3(29), 925, https://doi.org/10.21105/joss.00925