Computational Topology (Spring 2018) — Default Project

- **This is the default project. You could explore another real data set of your choice. Discuss options with me.**
- You **must email your report** as a **PDF file** to bkrishna@math.wsu.edu.
- You are welcome to send any code you create along with you submission. But you must explain all the details of how you went about implementing each step of the computation in your report.
- Your file name should identify you. For instance, if you are Eric Cartman, you should name your file EricCartman_Project.pdf. **Please start your name in this format, with the first letter in each subname capitalized**. If you want to add more bits to the title, e.g., Math574, you could name it EricCartman_Math574_Project.pdf, for instance. Also, please avoid white spaces in the file name :-).
- **This project is due by 5 PM on Friday, May 4**.

The goal of this project is for you to repeat the first part of the TDA pipeline used in the paper Topological Features In Cancer Gene Expression Data (arXiv:1410.3198). We analyzed gene expression data for five different cancers using persistent homology, and identified loops (i.e., non-trivial first homology) in all of them. We then checked the biomedical literature to see if the genes that formed these loops were implicated in the corresponding cancer. A majority of the loop-forming genes were reported to be implicated in their respective cancers in each data set.

Typical gene expression data sets contain expression of tens of thousands of genes for 10s or 100s of patients. As such these data sets are *high-dimensional* when viewed in the default sense. Instead, we *dualized* the data, i.e., looked at genes are points in the patients-space (in simpler words, we used the transpose of the data matrix). We then build witness complexes for increasing numbers of landmarks, and analyzed their persistent homology. In each data set, we found non-trivial first homology—1 or 2 highly persistent holes. We then identified the genes that formed these loops, and checked the literature for their relevance in the cancer of interest.

The Project

Repeat the first part of the analysis on at least **three** different gene expression data sets taken from Gene Expression Omnibus (available at https://www.ncbi.nlm.nih.gov/sites/GDSbrowser).

Use JavaPlex, Dionysus, or another similar tool for the persistent homology analysis.

Project Report: Explain clearly the assumptions you make about the data, as well as the steps you took to preprocess the same. Discuss the choices of number of landmarks (if using witness complexes), or the related parameters for building VR complexes. Also describe any computational analysis you could do to ascertain the statistical significance of your findings.

Limit your report to six pages.

You could work on a different topic for the project. Discuss the possibilities with me.