Computational Topology (Spring 2024): Default Project

- **This is the default project. You are encouraged to explore other real datasets of your choice for analysis using Computational Topology techniques, broadly defined. Discuss options with me.**
- You should submit a project report in a PDF file whose name should identify you in this manner. If you are Darryl Weathers, you should name your submission DarrylWeathers_Project.pdf. **Please start your name in this format. If you want to add details to the title, you could name it DarrylWeathers_Math529_Project.pdf, for instance. Please avoid white spaces in the file name :-).**
- If you want to **send additional files**, e.g., containing your code, then you should **include all files in a compressed folder**, e.g., zipped or tar-gzipped, named in the same manner as the main PDF file, i.e., **DarrylWeathers_Project.zip**. You could include the PDF file inside this zipped folder, along with the other files.
- **This project is due by by 11:59 PM on Friday, May 3**.

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, and are **encouraged to**, explore other topics. Discuss options with me.