Low Dose (LD) Radiation Biology (RadBio)

Carlos Deleon, Computer Science Department, Suffolk County Community College, Selden, NY 11784
Angelica Vanegas, Chemistry and Biology Department, Middlesex College, Edison, NJ 08837
Shinjae Yoo, Computational Science Initiate, Brookhaven National Laboratory, Upton, NY 11973

Abstract

Ionizing radiation was discovered in 1895, when Wilhelm Roentgen took an x-ray image of his wife's hand, since then radiation has been studied, including side effects. Since 1999 the US government has offered funding and resources for the study of low-dose radiation effects in the human population, especially to the Department of Energy Laboratories. Ionizing radiation can go through people's skin and be absorbed by tissue, this might cause harmful effects in the cell tissues, such as DNA damage, cancer, apoptosis of cells, alteration of gene expressions, and other effects. There are many ways how to identify them and visualize the change in cells once radiation is absorbed. Our research will focus on finding an efficient and effective way of recognizing the change in the radiated cell through machine learning visualization techniques.

Materials

Gene Expression Omnibus (GEO)
- GEO site is a large public database from National Center of Biotechnology Information that provides gene expression profiling information.
- Genomic data obtained from this public database

Python Libraries and Packages
- Pandas, NumPy, scikit-learn, Matplotlib, Seaborn
- Geo Parse

Google Colab
- Online Python programming platform
- Parsing data

Jupyter Notebooks
- Runs code locally (in PC), RAM, GPU
- PCA
- T-SNE
- Autoencoders

Methods

- Obtained from GEO site is our data GSE2109 the human genome atlas, and GSE43151 as the low dose ionizing radiation data.
- Probe ID to Gene ID
- Remove NaN (Not a Number) and duplicate values
- PCA color coded based on tissue type on annotation of GPL
- PCA on Joined data based on common Gene ID in both sets
- T-SNE of both data sets color coded on cell type.

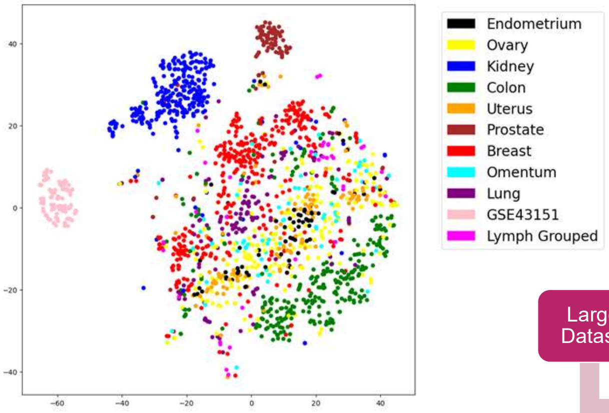

Figure 2: t- SNE of the most abundant cell types in GSE2109 joined with GSE43151 Lymph cells

No export control

Acknowledgement

- We would like to thank Brookhaven National Laboratory and our mentor Shinjae Yoo for selecting us for this opportunity to collaborate in this research. This project would have not been possible without our mentor's support and guidance, and his help for our learning development on research and critical thinking skills. We were exposed to many machine learning visualization methods, such as PCA, Autoencoder, t-SNE, and we will treasure these techniques for the rest of our careers. This project was supported in part by the U.S. Department of Energy, Office of Science, Office of Workforce Development for Teachers and Scientists (WDTS) under the Community College Internships Program (CCI).

References

- He, F., Yoo, S., Wang, D., Kumari, S., Gerstein, M., Ware, D. and Maslov, S. (2016), Large-scale atlas of microarray data reveals the distinct expression landscape of different tissues in Arabidopsis. Plant J, 86: 472-480. https://doi.org/10.1111/tpj.13175.
- Nosel I, Vaurijoux A, Barquinero JF, Gruel G. (Jul 12, 2013). Characterization of gene expression profiles at low and very low doses of ionizing radiation. *DNA Repair (Amst)* . (7):508-17. PMID: 23683873.
- Expression Project for Oncology (expO), geo, V1. (2005). https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2109.
- The pandas development team.(Feb, 2020).pandas-dev/**pandas**: Pandas.Zenodo.latest.10.5281/zenodo.3509134.**https**://doi.org/10.5281/zenodo.3509134.
- Pedregosa, F. **and** Varoquaux, G. **and** Gramfort, A. **and** Michel, V. **and** Thirion, B. **and** Grisel, O. **and** Blondel, M. **and** Prettenhofer, P. **and** Weiss, R. **and** Dubourg, V. **and** Vanderplas, J. **and** Passos, A. **and** Cournapeau, D. **and** Brucher, M. **and** Perrot, M. **and** Duchesnay. E. (2011). Scikit-learn: Machine Learning **in** Python . Journal of Machine Learning Research. Volume12. pg 2825--2830.
- Hunter, J. D. (2007). Matplotlib: A 2D graphics environment.Computing in Science \& Engineering. V {9} .N {3}. Pg {90--95}. . IEEE COMPUTER SOC. {10.1109/MCSE.2007.55}.
- Van der Maaten, L. & Hinton, G. (2008). Visualizing Data using t-SNE . *Journal of Machine Learning Research*, 9, 2579--2605.

Introduction

This Research focuses on visualization and analysis of radiation and their effect on healthy cells using machine learning techniques. The technique used was Principal Component Analysis (PCA), an effective linear graph approach. We chose PCA because it was a great place to start to understand our data, and it is well known for being a good technique to project a large set of data into a 2D space. Our data was obtained via the Gene Expression Omnibus (GEO) a large public database for genomic data. Our objective is to find a pattern in our PCA graph to be able to identify the different classifications of tissues based on clusters and observe where our radiation data falls within these clusters. We also use t-distributed stochastic neighbor embedding (t-SNE) to find non-linear patterns in our data which PCA can not normally find. It helps cluster different patterns found in the data while also displaying high dimensional data into a low dimensional space. Our final goal of this project is to have data generation to help scientist fill in missing gaps in data accurately which can lead to more analysis possibilities. Some ways this can be done is with Autoencoders or interpolation. This can help save time and money for scientists who need completed data or test points in bewteen a range.

Outcome

PCA Pattern: GSE2109 contained a cluster of patterns based on tissue type.
PC Percentages: The highest percentage in GSE2109 that could be explained in the PC percentage was 8% when matched with same genes common in GSE43151.There is nonlinear patterns which is present in T-SNE since its clusters are closer together.
Lymphocytes in PCA: The Lymphocyte TCD4 data from the radiation data compared to GSE2109, are more clustered in the middle of the graph. This could be because Lymph has different types of Lymphocyte in it when GSE43151 could be just one type of Lymphocyte cell.
T-SNE: Our graph shows clear clusters in our data show gene expression is consistent with cell type and can be used to determine it, but it also shows that batch effect is affecting our results for samples from other datasets.

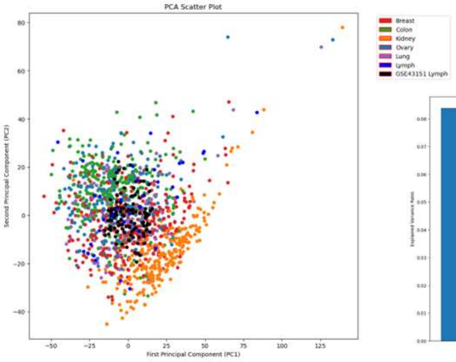

Figure 1: Joined data in PCA of data GSE2109 and GSE43151

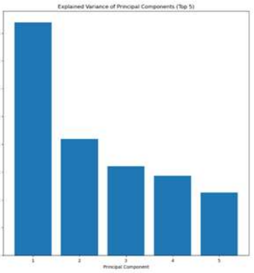

Figure 3: PC Percentage of 1-5 of GSE2109 with matching genes to GSE43151

Discussion

Gene Discussion

The PCA of joined data was accomplish by obtaining the Gene ID both datasets had in common. This is successful but only downside is we may be losing some important information

Batch Effect

There might be a non-biological factors affecting our data called batch effect due to datasets being from different laboratories and their way of data collection approach might be different between the two locations causing variations that effect analysis

Data size

- 121 sample size GSE 43151 and 2158 sample size in GSE2109
- Cleaned data for PCA had 15290 genes for size GSE 43151 and 51276 genes in GSE2109.
- 6021 common genes shared in both datasets.

Next steps

Larger Dataset →
- GSE2109 only has only 2158 samples but a new dataset E-MTAB-3732 on ArrayExpress has 27887 samples which allows us to have more data to work and test with. Due to its size, we will have to use the SDCC at BNL to handle this data thanks to BNL High Performance Computing.

Autoencoder / Visualization →
- Autoencoders are a way of taking data and reducing its features and its dimensionality to what we call a latent space
- Heatmaps can also be used to visualize patterns bewteen pairs seeing how it correlates, it can be another way to test if gene patterns are consistent bewteen different datasets.

Synthetic Data →
- With Autoencoders and other techniques we can now generate data
- Can be used to fill in caps of missing data or predict how sick a cell is.

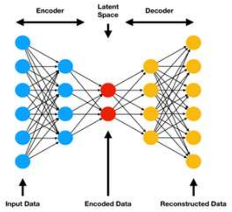

Figure 4: Autoencoder

Conclusion

- Data contains batch effect making our data not as reliable until it is removed
- Our data does have non-linear patterns
- Visualizations techniques that effectively graph data and show cell type
- This project once finished will help detect anomalies and changes in cells due to radiation or disease and even to predict the radiation effects

SUMMARY
- Parse data
- Analysis
- Visualizing
- Batch effect

PROJECT IMPACT
- Detect cancer in cells
- Predict cell health state

NEW SKILLS
- Machine learning visualization
- GEO database knowledge
- Python libraries

PROFESSIONAL GROWTH
- Teamwork skills
- Programming skills
- Problem-solving skills
- Research and development

www.bnl.gov