Apr 27th, 3:00 PM - 4:00 PM

Statistical Clustering of Glioblastoma Multiforme for Graph Theory Analysis

Jed Syrenne
jedsyrenne@gmail.com

Statistical Clustering of Glioblastoma Multiforme for Graph Theory Analysis

Jed Syrenne, Dr. Mark Grimes

The University of Montana, Division of Biological Sciences, Missoula, MT 59812

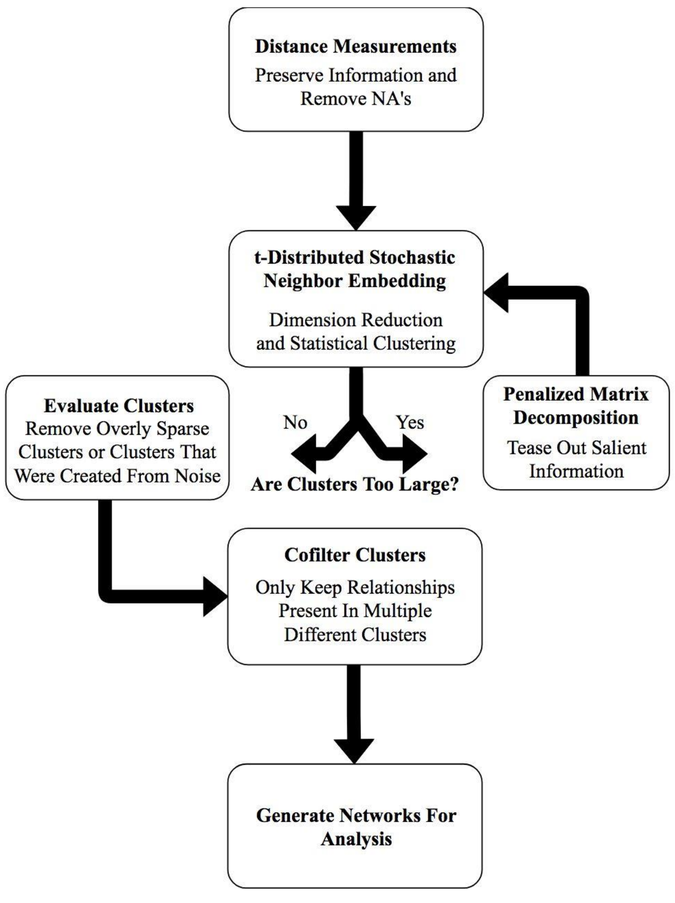
Introduction

Cells, whether single-celled or from a multicellular organism, constantly modify their behavior in response to their internal and external environments. These signals regulate differentiation, behavior, reproduction, movement, and programmed cell death. These signaling pathways are regulated via proteins which are post-translationally modified to carry messages. Understanding and analyzing signal transduction at a holistic level yields insights into the progress and treatment of innumerable diseases. However, there is a vast perplexity of proteins and modifications, making this an impossible task.

In statistical clustering, proteins that cluster together are likely to possess a functional relationship with each other. By statistically clustering and filtering proteomic data, networks can be created so that the vast perplexity of protein-protein interaction data can be understood and meaningfully analyzed. Dr. Mark Grimes has previously clustered and analyzed proteomics data from tandem mass-spectrometry experiments. Dr. Grimes utilized his clusterings to filter publicly available protein-protein interaction databases to create networks. These networks tease out the information from the obtuse databases.

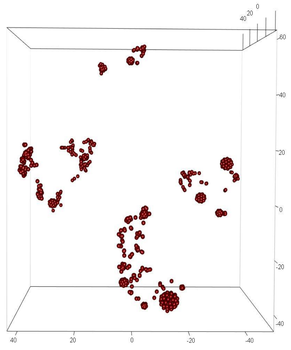
Here, we obtained publicly available glioblastoma proteomics data from PhosphoSitePlus and processed it using R and RStudios. These data were binary, showing only the presence of a protein instead of a ratio of the amounts of the protein. This produced a unique challenge for clustering and analysis, leading to a novel use of DINEOF for imputation in proteomics. Statistical clustering was performed and used to produce networks of glioblastoma protein-protein interactions for analysis.

Simplified Workflow

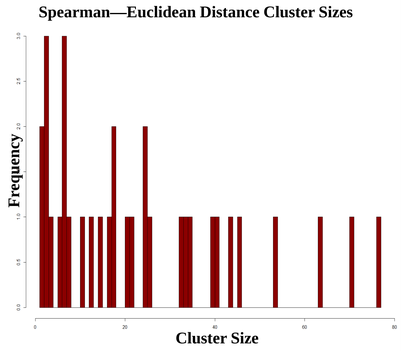


t-Stochastic Neighbor Embedding

A.



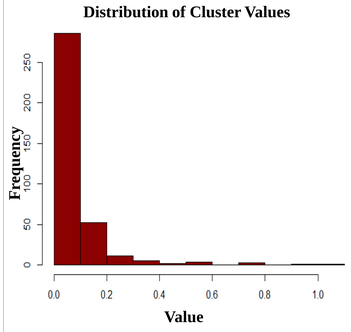
B.



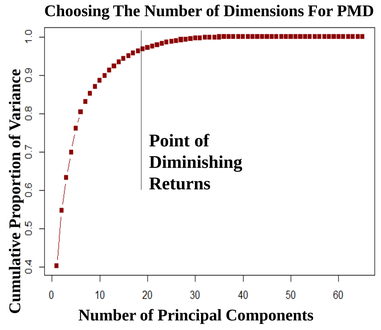
t-distributed Stochastic Neighbor Embedding, (t-SNE), is a machine learning algorithm that is useful in both clustering and dimensionality reduction. Here we utilized the algorithm for both clustering our data and reducing it from over 60 dimensions to 3 dimensions. **A.)** A 3-dimensional clustering of the data obtained from the Spearman Euclidean Distance, (SED). This distance measure is an equally weighted combination of both the Euclidean Distance and the Spearman Distance. Tight clusters are easily observed and are well separated. Each point in a cluster represents an individual protein. **B.)** A histogram of the number of proteins contained within each cluster obtained from the SED. Sizes range from 2 to 78 proteins. Clusters over the size of 60 proteins are considered oversized and require further refinement.

Refining Oversized Clusters

A.



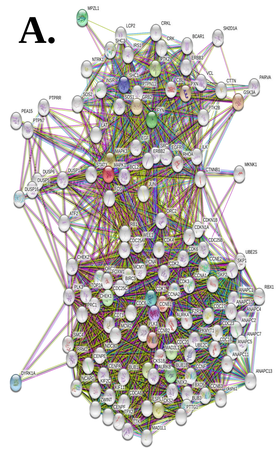
B.



A Penalized Matrix Decomposition, PMD, is used to break down the larger clusters and tease out the salient information. However, we have to choose how many dimensions we are going to reduce the cluster data down to. In order to accomplish this goal we impute NA's using a method known as DINEOF. This will preserve the very strong positive skew we see in **A.)** A distribution of the values found in a cluster. We then take the imputed matrix representing the cluster and use an algorithm similar to PMD, Principle Component Analysis. The principle components are analogous to the dimensions of PMD. **B.)** The cumulative number of principle components and the cumulative proportion of the variance of the data they explain. We choose the point of diminishing returns - where we have explained over 90 percent of the variance and taking an addition principle component does not explain an appreciable amount of variance. As principle components and dimensions of a penalized matrix decomposition are analogous, we take this number of principle components to be our number of dimensions.

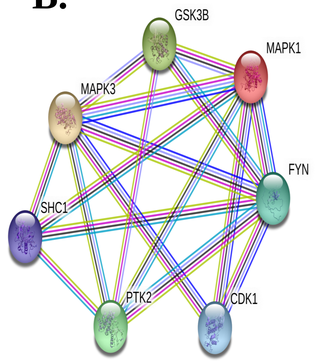
Protein-Protein Interaction Networks

A.



Filtering

B.



In **A.)** we see a "hairball" of protein-protein interactions from a public database, STRING. These databases contain high false positives and may over represent well studied proteins. We use the protein lists derived from our clusters to filter out many of the connections and to produce network **B**.), which can be analyzed and comprehended.

Conclusions and Future Directions

- **Binary data can be clustered if we collapse information by post-translational modification**
- **Continue producing and begin analyzing filtered networks**
- **Further test the validity of DINEOF**

References and Acknowledgements

- Dr. Mark Grimes – invaluable direction, leadership, mentoring
- UMCUR
- Davidson Honors College – Funding for poster printing

Grimes, Mark L. et al. "Wrangling Phosphoproteomic Data to Elucidate Cancer Signaling Pathways." Ed. Jorge Sans Burns. PLoS ONE 8.1 (2013): e52884. PMC. Web. 22 Apr. 2018.

Palacios-Moreno, Juan et al. "Neuroblastoma Tyrosine Kinase Signaling Networks Involve FYN and LYN in Endosomes and Lipid Rafts." Ed. Marco Punta. PLoS Computational Biology 11.4 (2015): e1004130. PMC. Web. 22 Apr. 2018.

Szklarczyk, Damian et al. "The STRING Database in 2017: Quality-Controlled Protein–protein Association Networks, Made Broadly Accessible." *Nucleic Acids Research* 45.Database issue (2017): D362–D368. *PMC*. Web. 22 Apr. 2018.

Hornbeck PV, Zhang B, Murray B, Kornhauser JM, Latham V, Skrzypek E PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Res. 2015 43:D512-20.