Unsupervised Learning Techniques In Modularizing Gene Regulatory Networks

EN.600.438 Computational Genomics: Data Analysis

Richard Chen, Steven Chen

Introduction

Gene regulatory networks (GRNs) have huge potential in modeling the kinetics of cell state transitions, and discovering relationships between genes and proteins. Various network inference methods such as LASSO and Bayesian networks are widely used for GRN reconstruction, however there does not exist a single inference method that can scale well with large networks. Graphical LASSO (GLASSO), for example, has one of the fastest run-times of all inference methods, but suffers from low precision and recall. Bayesian networks on the other hand, are accurate in modeling small networks, but have exponential time complexity as the network size increases. A comparative study by Marbach *et al.* 2014 tested and validated over 30 inference methods on the DREAM5 challenge data set, in which all individual inference methods performed poorly. In the landscape of GRN reconstruction methods, there exists a strong need for an inference method that preserves run time and precision on large networks. In this project, we aim to explore new techniques that might improve GRN reconstruction speed and performance.

By the end of our project, we will have answered these questions:

1. Can we reduce the search space of GRN reconstruction using other network machine learning algorithms?
2. How can we verify that these methods preserve the biological integrity of the network?
3. What potential next steps can we take using the reduced search space for GRN reconstruction?

Related Work

Many attempts have been made in the field of gene regulatory networks that sought to improving precision and accuracy in various network inference methods, some of which were outlined in Marbach *et al.*. In 2003, Segal *et al.* introduced a community network based clustering methodology called module networks, which was then applied to genome-wide expression studies in Xu *et al.* in 2004. In 2014, Marbach *et al.* evaluated various different network inference methods, including module networks introduced in Segal *et al.*, and concluded that community based algorithms like module networks outperformed individual inference methods like Bayesian network inference or GLASSO. Our project sought to integrate the findings from Segal and Marbach, and expand upon and verify the community-based concept behind module networks by using other unsupervised machine learning algorithms. We hope that this will both improve precision, and reduce computational run time by reducing the search space of network inference methods while preserving biological significance.

Data

Our data is taken from the DREAM5 simulated mRNA expression data used to assess various network inference methods. These networks are based on characteristics of a well-studied system, the *E. coli* K-12 strain. The data was retrieved from the Marbach *et al.* comparative study, where a more detailed description of that data can be found. Originally, the *E. coli* system had expression data with 4511 genes over 805 samples. However, because computational run time was a significant concern for the project, we decided to reduce the size of this system to reflect only significant relationships between genes based on the gold standard network provided for the system. In other words, we removed any gene that did not have a connected edge in the gold standard network.

Data Acquisition and Normalization

From Marbach *et al.*:

> A compendium of microarray data was compiled for *E. coli*, where all chips are the same Affymetrix platform, the *E. coli* Antisense Genome Array. Chips were downloaded from GEO (Platform ID: GPL199 and GPL). In total, 805 chips with available raw data Affymetrix files (.CEL files) were compiled for *E. coli*.
>
> Microarray normalization was done using Robust Multi-chip Averaging (RMA) through the soft-

ware RMAExpress. All 805 chips were uploaded into RMAExpress and normalization was done as one batch. All arrays were background adjusted, quantile normalized, and probesets were summarized using median polish. Normalized data was exported as log-transformed expression values. Mapping of Affymetrix probeset ids to gene ids was done using the library files made available from Affymetrix. Control probesets and probesets that did not map unambiguously to one gene were removed. Lastly, if multiple probesets mapped to a single gene, then expression values were averaged within each chip.

Methods

In Young *et al.* 2015, Young found that Bayesian networks had the highest precision value and area under the curve (AUC) statistic amongst popular inference methods such as LASSO, ARACNE, and CLR on small networks, however, its performance dropped off on larger networks. To reduce the search space of calculating prior and posterior probabilities, we will split our expression data using clustering algorithms such as gaussian mixture models (GMMs), k-means, spectral clustering, Fast Louvain, and Clauset-Newman-Moore (CNM), and then validate the resulting communities by calculating frequency of significant gene ontology (GO) pathways within each cluster. We hypothesize that given a large gene expression matrix, our community detection algorithms would cluster genes that share the same pathway. We assume that our community detection algorithms would be able to detect overlaps for genes that exist in more than one pathway.

Clustering

Our main goal in this project is to cluster the data using the algorithms we mentioned above - GMM, k-means, Spectral Clustering, CNM, and Fast Louvain. Since we aren't certain about the data characteristics, we use algorithms that cluster based on different features in the data.

We used k-means and GMMs to cluster to emphasize data compactness in terms of euclidean distances between means, spectral clustering to emphasize connectivity and affinities between related data, and large network clustering methods like Fast Louvain and CNM to extract communities from initialized networks from the data.

The k-means and GMM clustering algorithms only required the expression data matrix to cluster because of their dependencies on euclidean distance, but Fast Louvain, Spectral Clustering, and CNM were network-based, and required an initial adjacency matrix. In these algorithms, we first used two network inference algorithms, GLASSO and Weighted Correlation Network Analysis (WGCNA) in order to construct an initial adjacency matrix from the data.

To determine the optimal number of clusters for each method, we performed the algorithms onto a range of numbers of clusters, and chose the number with the highest Bayesian Information Criterion (BIC) score. As a result, a unique optimal number of clusters is obtained for each clustering method. Additionally, we obtained the distribution of cluster sizes for each clustering algorithm to empirically evaluate whether or not the algorithm was outputting reasonable cluster sizes and cluster distributions.

To evaluate performance of our different clustering methods, we assigned each gene in each distribution of clusters to their corresponding GO pathway category, and determined the frequencies at which the same GO pathways were being represented within each individual cluster. From this, we used a hyper-geometric test to determine whether or not the cluster contained a significant GO pathway.

Results

Note: All figures are in the appendix section of the report.

The optimal number of modules, or clusters, for each algorithm we ran is shown in Figure 1. The number was evaluated using the highest BIC score for a range of numbers, an example of which for K-means is shown in Figure 2.

The cluster size distributions for each method is shown in Figures 3 through 10. Empirically, the distribution for the Fast Louvain clustering for both GLASSO and WGCNA performed poorest. Most of the genes in the expression data were put into one cluster, which doesn't accurately represent the separate biological communities between highly correlated genes. The other network-based clustering algorithms performed slightly better in terms of capturing communities, but GMMs and K-means clustering performed the best. These algorithms both had a smaller optimal number of clusters (see Figure 1), and exhibited a more evenly distributed spread than any network-based algorithm that we tested.

This inconsistency may be attributed to the properties of the initial network inference models we made. As seen in Figures 11 and 12, the node eccentricity for the entire network is rather low. This means that the inferred networks are disconnected, and as a result very short pathways are formed. This attributes to the inaccuracies and low precisions of the network inference models as outlined in Marbach *et al.*, and may contribute to the low distribution spread of the subsequent clustering methods as described in the previous section.

Figures 14, 15, 16, and 17 show the frequency of significant GO pathways in K-means, GMM, and GLASSO-based and WGCNA-based Spectral Clustering. These results show that the distribution of significant GO pathways in both K-means and GMM clustering is sparse, and seem to contradict the results previously observed where these clustering methods produced a more evenly spread cluster size distribution. In comparison, the spectral clustering algorithm performed better in that the distribution of GO pathways are more evenly spread across all clusters. The more even spread indicates that each cluster is more specific to one particular biological function, or GO pathway, which was the ultimate goal in preserving the biological integrity of the network.

Conclusions

The project produced some interesting results. We concluded that though the cluster size distribution for euclidean distance based clustering methods like k-means and GMM seem to be more representative of the true biological network for a genomic system, there was evidence shown that the network-based clustering algorithms actually represented these networks better when validated with the GO pathway categories. This conclusion may infer that the preferred unsupervised learning technique for gene expression data may be network-based.

Though this project produced interesting results, there are many more future steps that could be taken to further validate the claims we made, and further reach the ultimate goal of reducing computational run time and increasing accuracy and precision for network inference methods. Though we were limited by time and computational power for this project, future projects could look to use the same algorithms on different well-studied biological data sets, and with enough computational power, perform different types of network inference methods onto the individual clusters obtained from the algorithms in this project.

References

http://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-8-47

http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0029348#references

http://www.ncbi.nlm.nih.gov/pubmed/22796662

http://arxiv.org/pdf/1110.5813v4.pdf

https://www.bioconductor.org/packages/3.3/data/experiment/vignettes/DREAM4/inst/doc/DREAM4.pdf

http://www.ncbi.nlm.nih.gov/pubmed/12740579

http://onlinelibrary.wiley.com/doi/10.1016/j.febslet.2004.11.019/full

http://ece-research.unm.edu/ifis/papers/community-moore.pdf

Course project statements

This project was not related to previous research or other outside projects. Richard Chen wrote the code script, and Steven Chen wrote the write-up.

Appendix

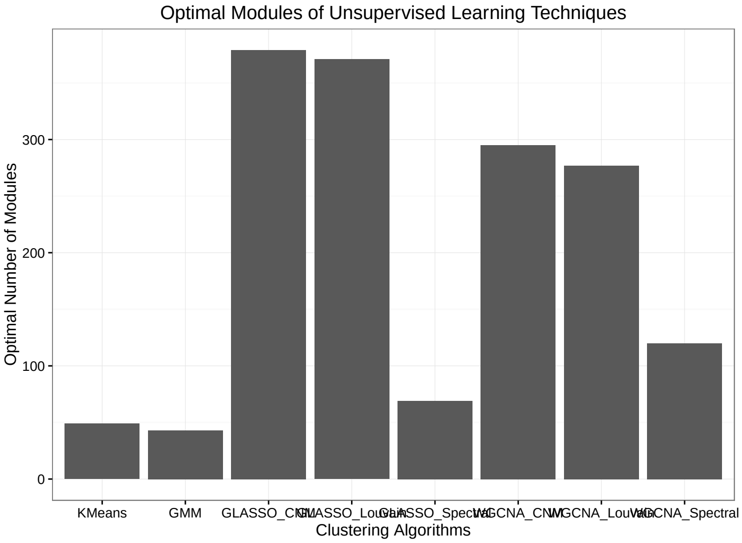

Figure 1

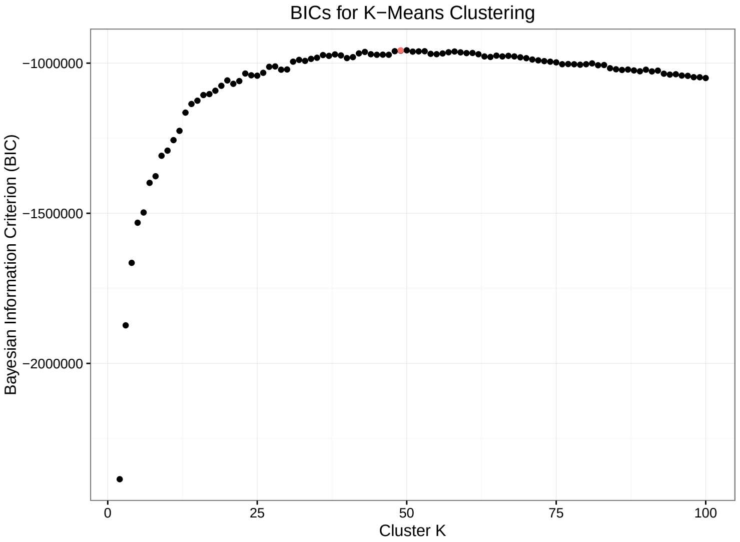

Figure 2

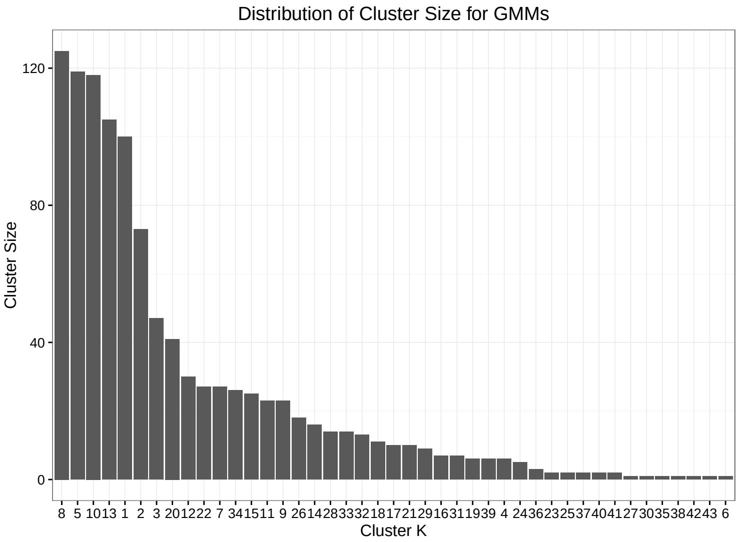

Figure 3

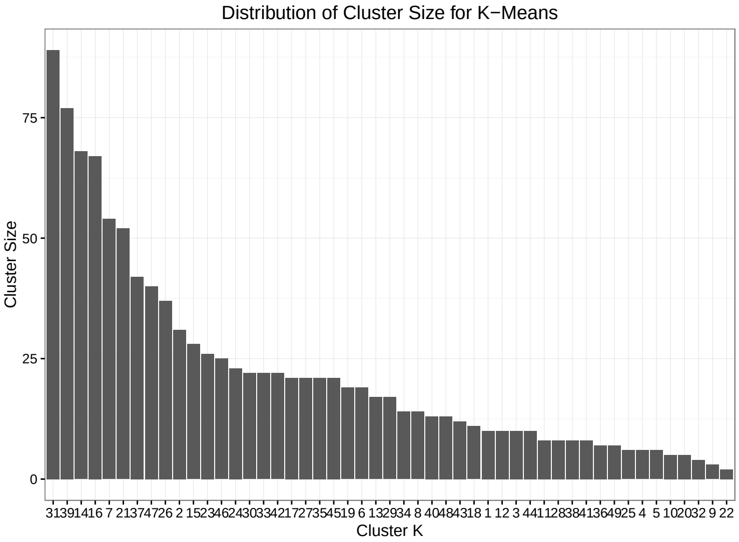

Figure 4

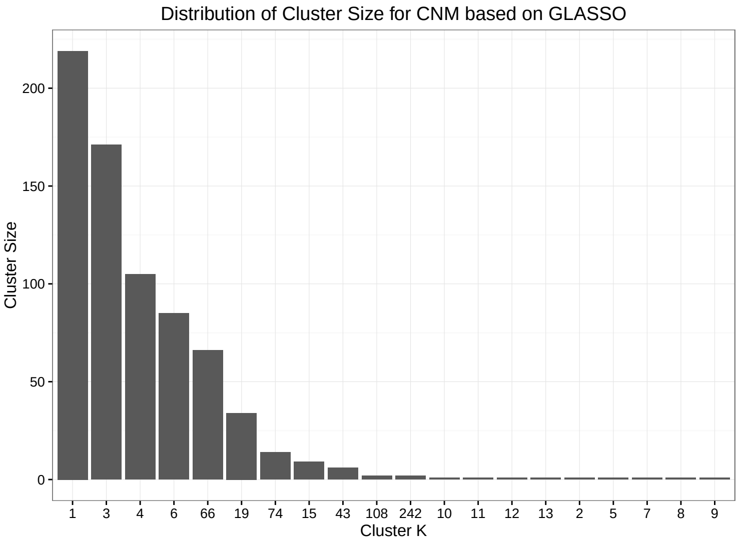

Figure 5

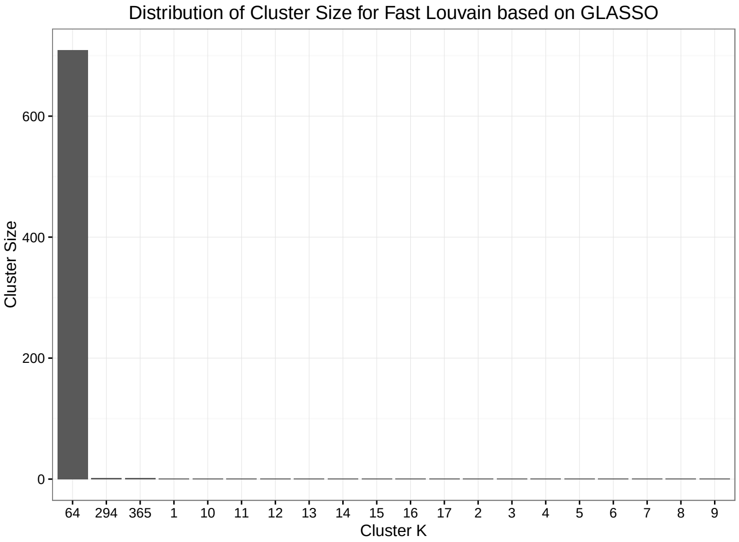

Figure 6

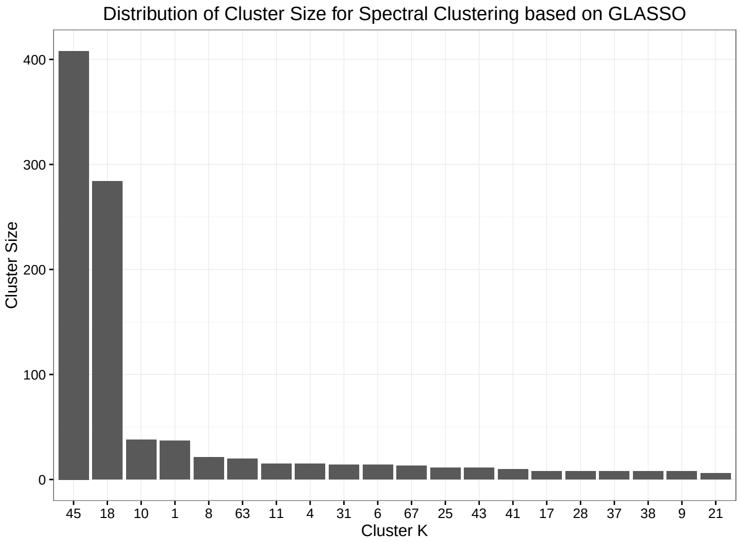

Figure 7

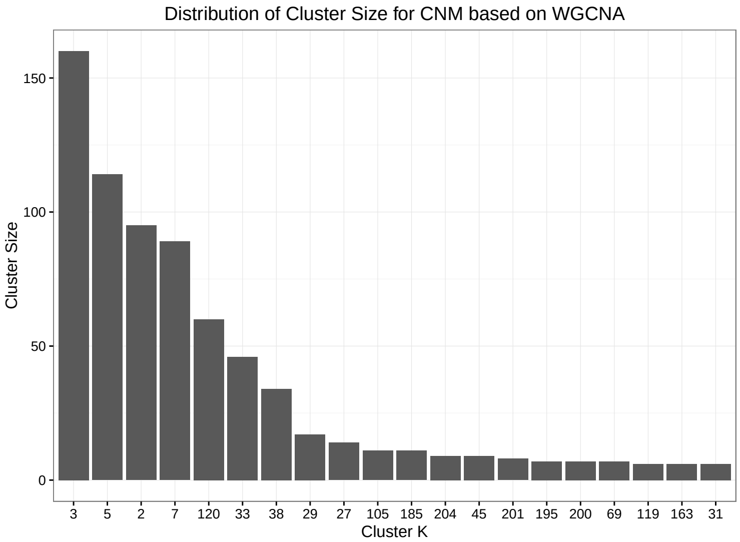

Figure 8

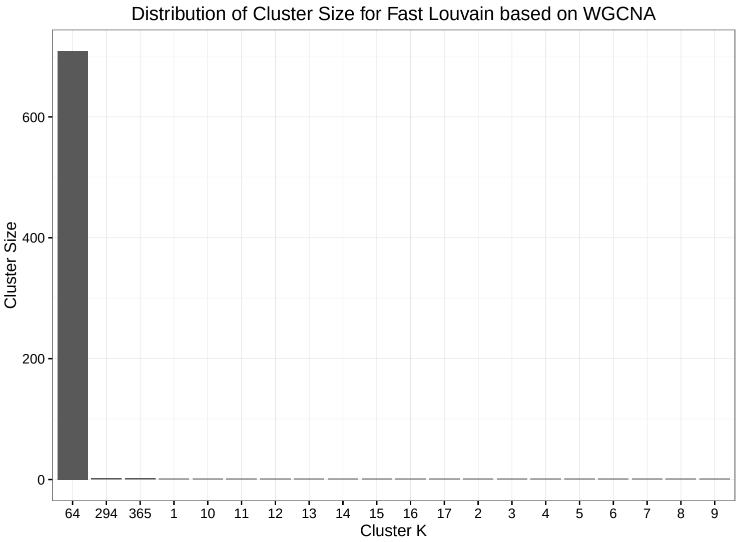

Figure 9

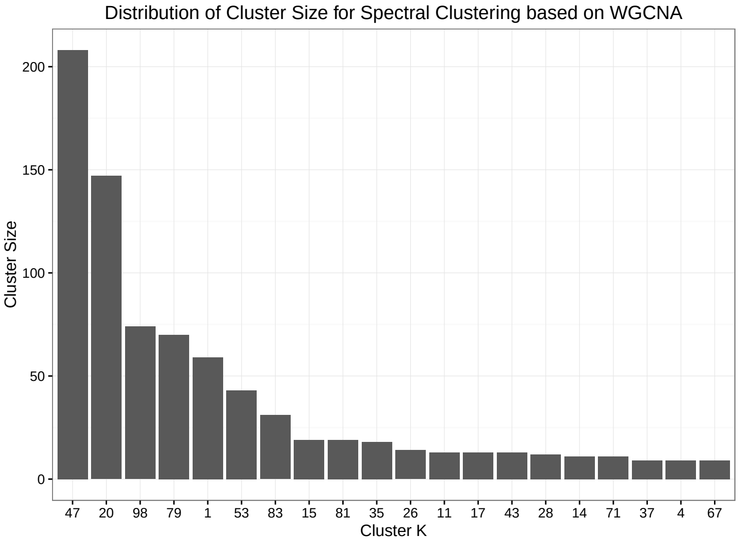

Figure 10

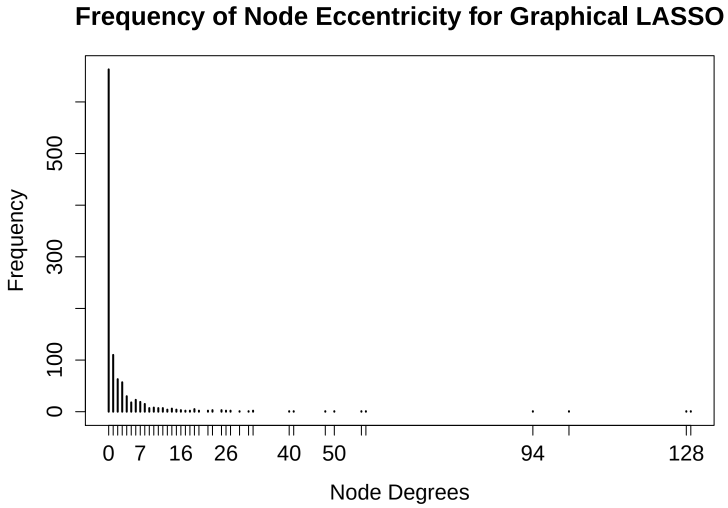

Figure 11

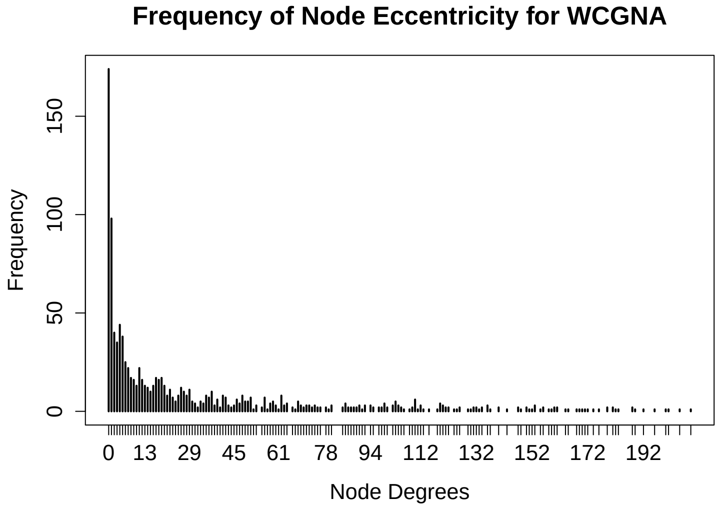

Figure 12

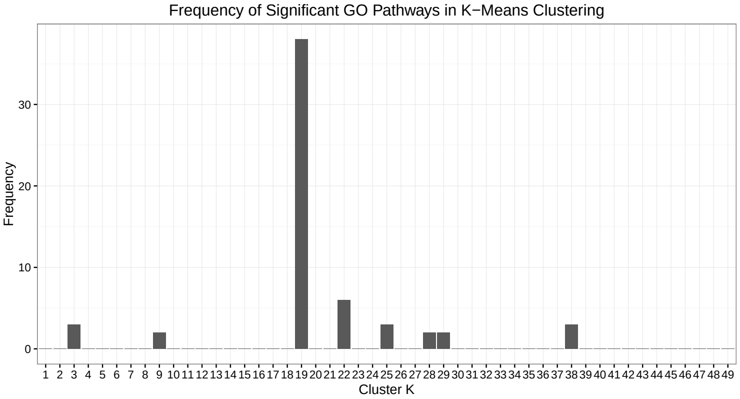

Figure 13

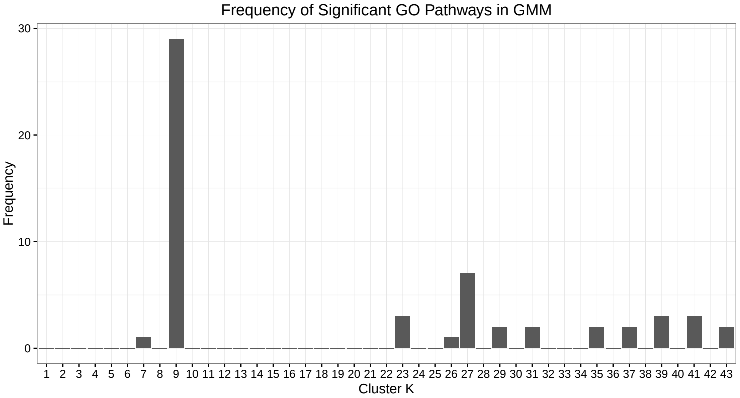

Figure 14

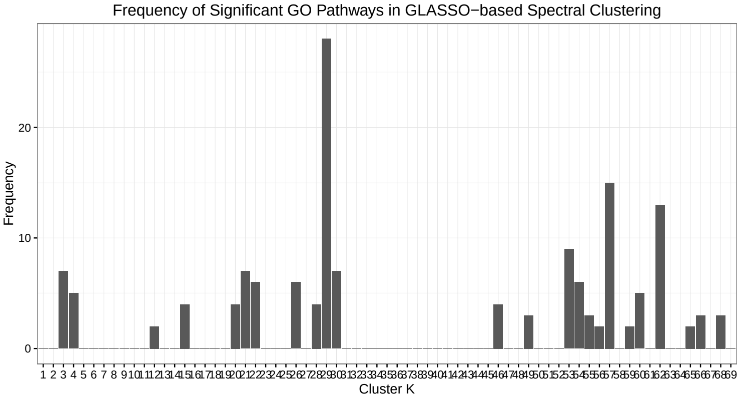

Figure 15

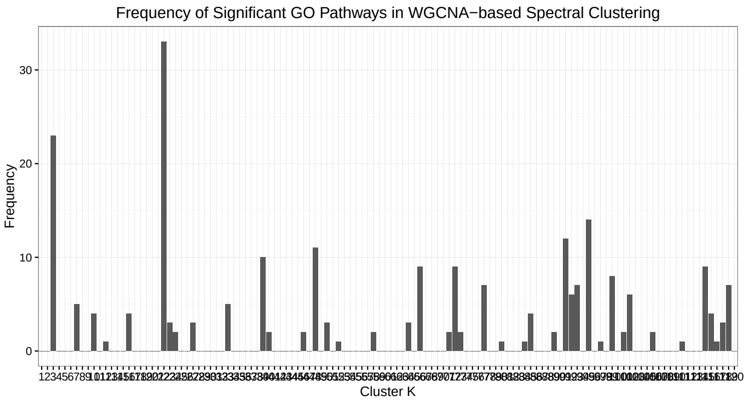
Frequency of Significant GO Pathways in WGCNA−based Spectral Clustering
Frequency
30
20
10
0
Cluster K

Figure 16