Leveraging Stochastic Block Model for Uncovering Genomic Associations in Multilayer Networks

Reddy Rani Vangimalla*, Ankur Pachauri

Department of Computer Science & Information Systems, Birla Institute of Technology and Science, Pilani, India

Email address:

reddy.rani@pilani.bits-pilani.ac.in (Reddy Rani Vangimalla), ankur.pachauri@pilani.bits-pilani.ac.in (Ankur Pachauri)

*Corresponding author

Abstract

Integration of multiomics data for gene-gene association is widely performed in genotype-phenotype studies. Genes associated with disease pathways are crucial for the design of precision medicine, patient subtyping, and drug prediction. This study can extend to various diseases to derive a disease- specific network and expand to create a disease-phenome and disease-genome bipartite network, known as a Diseasome. We propose integrating multilayer networks of a disease and identifying communities of co-associated genes using the Stochastic Block Model (SBM) approach. SBM was originally used for generating random probabilistic graphs with communities or clusters embedded within. Some researchers coin the term planted clusters as the number of clusters or communities has to be fixed, beforehand. The heterogeneity between the connections is modelled in SBM using the probabilistic approach to understand the blocks or clusters in the network. This model uses probabilities of connections within the blocks and between the blocks. The method generates a symmetric probability matrix with dimensions K ×K matrix P of K communities, where diagonal values represent the probability of nodes connecting within the same community, and off-diagonal values represent the probability of nodes connecting between the communities. These communities represent the latent structure of a network and serve as a testbed with the benchmark results for multilevel/multilayer networks and dynamic networks. We propose using SBM for an undirected graph where sets of genes, represented as nodes of the network, which will be divided into clusters. The clusters generated, using this model, will be structurally similar, in our work, we will be using this property for understanding genomic associations. To identify the number of communities (K), we use a voting process for each value of K, starting from 2 clusters. We measure quantitative scores, including Modularity, Normalized Mutual Information, Silhouette Score, Dunn Index, Jaccard Similarity, Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC), and the elbow point, to determine the best-fit value for K. We further analyze the genes of each community and validate their biological significance using gene enrichment analysis. To extend our understanding, we verify the pathways of genes in each community by considering WikiPathways, KEGG pathways, and Reactome pathways. Finally, we compare SBM with modularity maximization procedures and consensus-based community detection methods to draw our conclusions.

Keywords

Multiomics Network Generation, Stochastic Block Modeling, Disease Gene Associations, Diseasome - Bipartite Network