Using Graph Neural Networks for Identifying Overlapping Modules in Gene Co-expression Networks

Camila Riccio, Jorge Finke, and Camilo Rocha

Pontificia Universidad Javeriana, Cali, Colombia
{camila.riccio,jfinke,camilo.rocha}@javerianacali.edu.co

A gene co-expression network is an undirected *graph* $G = (V, E)$ where each *vertex* in V corresponds to a gene, and a pair of vertices (u, v) is an *edge* in E if the genes represented by u and v have at least one expression pattern in common. Modules represent basic structures for understanding the network organization. More specifically, modules are defined as groups of genes with similar expression profiles, which also tend to be functionally related and co-regulated.

Several approaches for module detection in gene co-expression networks have been proposed [4]; they mainly focus on the discovery of non-overlapping modules. However, gene co-expression modules are likely to overlap due to the multiple regulatory domains a gene can be part of. Towards this direction, Neural Overlapping Community Detection (NOCD) has been recently proposed [5] as an alternative to the discovery of modules with genes in common. It is based on a neural network that exploits the structure of graphs. Neural networks are a collection of connected nodes called *neurons* aggregated into layers. Data travel from the input to the output layer, after going through one or multiple hidden layers, where different data transformations are performed [6]. The NOCD model takes as input a network adjacency matrix and/or an attribute matrix. It outputs an affiliation matrix assigning nodes into a fixed number of possibly overlapping modules. Because of their ability to reproduce and model nonlinear processes, neural networks have found applications in many disciplines. NOCD is a pioneer in the detection of overlapping modules using neural networks. Thus, applying NOCD in co-expression networks is a novel and promising approach.

This work presents preliminary work on the use of NOCD for discovering modules, with possible overlaps, in a co-expression network of rice under salt stress. The network is built from data available on GEO [2] under the accession number GSE98455. Our main goal is to evaluate the performance of the NOCD model detecting meaningful overlapping modules in the context of co-expression networks and to compare the outcome to the one generated by the real gene affiliation matrix. In particular, gene ontology information is used to build such a matrix. The agreement between true and detected communities is quantified using the overlapping normalized mutual information metric [3]. Additionally, NOCD is compared with Hierarchical Link Clustering (HLC) [1], another overlapping module detection technique that can be used in co-expression network analysis. The quality of the communities found with both techniques is evaluated using unsupervised metrics such as coverage, density, conductance, and clustering coefficient.

Finally, it is important to note that NOCD is highly scalable. Thus incorporating this novel clustering technique enables us to handle large graphs, which is a key requirement for finding more meaningful gene modules.

References

1. Ahn, Y.Y., Bagrow, J.P., Lehmann, S.: Link communities reveal multiscale complexity in networks. Nature **466**(7307), 761–764 (2010)
2. Clough, E., Barrett, T.: The gene expression omnibus database. In: Statistical Genomics, Methods in Molecular Biology, vol. 1418, pp. 93–110. Humana Press, New York, NY (2016)
3. McDaid, A.F., Greene, D., Hurley, N.: Normalized mutual information to evaluate overlapping community finding algorithms. arXiv preprint arXiv:1110.2515 (2011)
4. Saelens, W., Cannoodt, R., Saeys, Y.: A comprehensive evaluation of module detection methods for gene expression data. Nature Communications **9**(1), 1–12 (2018)
5. Shchur, O., Günnemann, S.: Overlapping community detection with graph neural networks. arXiv preprint arXiv:1909.12201 (2019)
6. Wang, S.C.: Artificial neural network. In: Interdisciplinary computing in java programming, pp. 81–100. Springer (2003)