05 Jul 2018

Analysis Of Grapevine Gene Expression Data Using Node-based Resilience Clustering

Jeffrey Dale

John Matta

Susanne Howard

Gunes Ercal

et. al. For a complete list of authors, see https://scholarsmine.mst.edu/ele_comeng_facwork/4828

Recommended Citation

J. Dale et al., "Analysis Of Grapevine Gene Expression Data Using Node-based Resilience Clustering," *2018 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2018*, pp. 1 - 8, article no. 8404962, Institute of Electrical and Electronics Engineers, Jul 2018.
The definitive version is available at https://doi.org/10.1109/CIBCB.2018.8404962

Analysis of Grapevine Gene Expression Data using Node-Based Resilience Clustering

Jeffrey Dale
Computer Science Dept.
Missouri State University
Springfield, USA
dale2755@live.missouristate.edu

John Matta
Computer Science Dept.
Southern Illinois University Edwardsville
Edwardsville, USA
jmatta@siue.edu

Susanne Howard
College of Agriculture
Missouri State University
Mountain Grove, USA
susannehoward@missouristate.edu

Gunes Ercal
Computer Science Dept.
Southern Illinois University Edwardsville
Edwardsville, USA
gercal@siue.edu

Wenping Qiu
College of Agriculture
Missouri State University
Mountain Grove, USA
wenpingqiu@missouristate.edu

Tayo Obafemi-Ajayi
Engineering Program
Missouri State University
Springfield, USA
tayoobafemiajayi@missouristate.edu

***Abstract*—Powdery mildew is the most economically important disease of cultivated grapevines worldwide. In the agricultural community, there is a great need for better understanding of the complex genetic basis of powdery mildew (PM) resistance by delineating possible gene biomarkers associated with the plants' defense mechanisms. Machine learning techniques can be applied to analysis of gene expression data to aid knowledge discovery of disease fighting genes. In this work, we apply a data-driven computational model, utilizing a graph-based clustering algorithm – Node-Based Resilience Clustering (NBR-Clust), to analyze grapevine gene expression data to identify possible gene biomarkers associated with powdery mildew disease defense mechanisms. We investigated two graph representations (geometric and kNN) on the mean differences of PM inoculated vs. mock inoculated gene expression values of Cabernet and Norton (PM disease resistant) species across 6 time points. By applying the contrarian approach, we hypothesized that smaller sized clusters will contain genes that do not follow general patterns, hence, could display distinct expression patterns of PM-induced transcripts across the time points that may insinuate biological relevance. We compared the smaller clusters obtained in Norton in contrast with the ones from Cabernet in terms of the genes that clustered in common between both (intersection of sets) as well as the differences of the sets. The results obtained demonstrate the usefulness of the geometric graphs for this domain application in contrast to the kNN graphs. Some genes that belong to biologically relevant pathways were identified that displayed differences in patterns across the time points between Norton and Cabernet species.**

***Index Terms*—plant disease resistance, genes, clustering, graph theory, resilience measures.**

I. Introduction

Machine learning (ML) is increasingly becoming a fundamental foundational approach for interdisciplinary research as intelligent data analytic systems span diverse applications beyond computer science and engineering. ML methodologies are useful in biological applications to characterize, process and integrate data to uncover useful information and actionable knowledge. ML encompasses different approaches of learning including supervised learning, unsupervised learning (also known as clustering) and reinforcement learning. Clustering techniques are particularly useful in domains where there is no ground truth or class information.

In this work, we investigate a data driven methodological framework using a novel graph-based method, node-based resilience clustering (NBR-Clust) [1], to infer meaningful biomarkers from heterogeneous noisy biological gene expression data. This study is a translational research that focuses on transforming concepts and theories into practical applications. Our domain application addresses disease resistance mechanisms of grapevine. The most economically important disease of cultivated grapevines worldwide is powdery mildew. In the agricultural community, there is a great need for better understanding of the complex genetic basis of powdery mildew resistance, and subsequently, genes associated with plants' defense mechanisms. Large-scale gene expression data coming from micro-array experiments provide new means to reveal fundamental cellular processes, investigate functions of genes, and understand relations and interactions among them. Our objective is to use this data to identify and characterize the set of genes responsible for powdery mildew resistance using novel graph-based clustering techniques.

Grapevine is one of the most cultivated fruit crops in the world [2], [3]. Powdery mildew (PM) is an economically important disease of grapevines that causes significant losses in yield and reduction in berry quality [3], [4]. To prevent this disease, fungicide is applied during production which translates to increased production costs. This also poses significant risk to the health of growers and of the environment pollution. Certain grapevine species have demonstrated a high degree of resistance to this disease, such as Missouri's state grape: Norton (*Vitis aestivalis*). However, Norton grapes are of low quality [4], [5] and higher quality grapes, such as Cabernet

This work was partially supported by Missouri State University Faculty Research Grant. Corresponding author: Tayo Obafemi-Ajayi.

978-1-5386-1399-3/18/$31.00 ©2018 IEEE

Sauvignon (*Vitis vinifera*), are highly susceptible to PM.

Given the critical need to identify the role of certain genes in PM disease resistance, some work [2], [4]–[7]) has been done in genomic analysis of grapevine species. Qiu et al. in [3] present a detailed review of current knowledge from multiple studies on the genetic basis of the PM disease. Gene expression studies has identified some genes that have shown differential expression levels between powdery mildew resistance in the wild Vitis species and the susceptible V. vinifera varieties [4]. For example, there is some evidence suggesting existence of a PEN1-mediated secretory pathway is an important component of pattern triggered immunity against PM in grapevine [3]. There is still a need for more targeted approaches to explore grapevine genes involved in PM susceptibility and pathways underlying this process.

We present a data driven methodological framework that analyzes a set of gene expression data from both Norton and Cabernet cultivars using graph-based clustering algorithm. The aim is to uncover and more clearly delineate disease resistance patterns and features that would increase our understanding of how plants fight against disease. We are interested in observing the differences or similarity in expression patterns of Cabernet vs. Norton across biologically relevant genes in response to the PM inoculation across different time points. The hypothesis is that there are certain genes associated with PM defense mechanisms in grapevines.

The remainder of this paper is organized as follows. In Section II, we present an overview of the NBR-Clust algorithm and graph types. In Section III, we describe the proposed framework for detection of significant set of genes that are could be associated with PM disease fighting mechanism. The results obtained are demonstrated and analyzed in Section IV. The conclusion and next steps are discussed in Section V.

II. Background

To provide a context for the graph-based ML approach presented in this work, we briefly describe our underlying node-based resilience clustering (NBR-Clust) algorithm, the fundamentals of graph representations considered, and the grapevine gene expression dataset. The source code of the NBR-Clust algorithm is available at http://www.cs.siue.edu/∼gercal/clustering/.

A. Node-Based Resilience Clustering

Node-based resilience clustering [1] is a graph-based clustering framework that utilizes the *critical attack sets* returned by graph theoretic resilience measures to cluster a graph. Every node-based resilience measure involves computation of a critical attack set of vertices S such that the removal of S results in a relatively significant disruption to the remaining network. Attack sets identify weaknesses or bottlenecks in graphs [8], and removal of the corresponding nodes breaks the graph into disjoint partitions. The NBR-Clust algorithm has been explored for five resilience measures [1]. Matta et al. in [1] discuss the different properties of the NBR-Clust algorithm based on each type of resilience measure. In this work, we apply NBR-Clust using the normalized integrity measure [9]. Normalized integrity is defined as:

$$I(G) = \min_{S \subset V} \left\{ \frac{|S| + C_{max}(V - S)}{|V|} \right\}, \tag{1}$$

where $G = (V, E)$ denotes a graph with a set of V vertices and E edges, S is the attack set which is a subset of V, and $C_{max}(V - S)$ is the largest connected component in $V - S$.

One of the main advantages of NBR-Clust is the ability to cluster in one step (i.e. not requiring multiple recursive iterations) where the number of clusters is not known *a priori*. Results from clustering using a resilience measure like integrity have been shown to indicate a natural number of clusters for a graph [1]. The experimental results in [1] demonstrated the usefulness of integrity-based clustering in finding the optimal number of clusters. Since the number of clusters sought in this work was unknown, NBR-Clust based on integrity was employed in this work. The NBR-Clust framework, as applied, consists of following three key steps.

NBR-Clust Framework

1. Convert data to a graph representation, G.
2. Compute integrity resilience measure ($I(G)$) and its corresponding attack set S.
3. Remove the attack set nodes S from the graph and output the resulting clusters. (Attack set nodes are not reassigned in this work).

Given the computation hardness of computing integrity [10], a heuristic method known as GreedyBC was utilized to estimate integrity in step 2 of the NBR-Clust framework. GreedyBC relies on graph-theoretic measure called *betweenness centrality*. Using GreedyBC, the highest betweenness nodes are repeatedly removed from the graph, and with each removal the integrity resilience measure $I(S, G)$ is computed at that configuration. The configuration with the lowest integrity score is taken to be $I(G)$ and nodes S removed to that point are considered to be the attack set. The GreedyBC algorithm based on integrity resilience measure $I(G)$ can be summarized as:

Greedy-BC Heuristic

1 $I_{min} = I(G)$, $S_{min} = \{\}$
2 repeat $|V|$ times
3 $\quad v = argmax_{v \in V} BC(V)$
4 $\quad G = G \setminus \{v\}$ and $S = S \cup \{v\}$
5 $\quad$ if $I(S, G) < I_{min}$ then
$\quad\quad I_{min} = I(S, G)$ and $S_{min} = S$
6 return S_{min}

B. Graph Representations

The study of applicable graph representations is fundamental to the success of graph-based clustering approaches. The first step of NBR-Clust is to convert the input data into a graph representation. For the same dataset, different graph representations could result in different edge weights, densities, numbers of connected components, and modularities.

TABLE I
PROPERTIES OF GEOMETRIC VS. KNN GRAPHS GENERATED

Cultivar	Threshold Power/radius	Nodes	Edges	Average Degree	Modularity
Cabernet	pow5	9,113	727,569	160	0.61
Cabernet	pow6	9,112	537,210	118	0.63
Cabernet	pow7	9,110	414,342	91	0.64
Cabernet	kNN-5	9,113	16,928	4	0.91
Cabernet	kNN-30	9,113	111,551	25	0.74
Norton	pow5	9,113	534,786	117	0.66
Norton	pow6	9,112	388,486	85	0.68
Norton	pow7	9,112	294,993	65	0.70
Norton	kNN-5	9,113	16,876	4	0.91
Norton	kNN-30	9,113	113,297	25	0.73

Selection of appropriate graph representation influences the subsequent results obtained. In this work, we investigate the effectiveness of two graph types: geometric and k-nearest neighbors (kNN) graphs.

1) Construction of Geometric Graphs: The Weighted Gene Correlation Network Analysis (WGCNA) R package [11] was employed to convert the grapevine gene expression data to desired geometric graphs. Each node in a geometric graph is connected to all nodes within a certain radius r, or distance. WCGNA returns an adjacency matrix whose entries are the correlations between the gene indicated by the row of the input data matrix and the gene indicated by the column of the input data matrix. All of the entries of the adjacency matrix are in $[-1, 1]$. WGCNA computes correlation using either Pearson or Biweight Mid-Correlation measure (bicor) [12]. We selected bicor as it has been demonstrated in literature to be more robust [13].

WGCNA requires specification of the thresholding power parameter t_p to determine the minimum level of correlation between genes required for inclusion in the graph. Let B denote the final adjacency matrix with the less-correlated genes filtered from the original adjacency matrix A. In general, for an entry a_{ij} of A and a soft thresholding power t_p, b_{ij} is given by

$$b_{ij} = \left\{ \begin{array}{ll} a_{ij}, & \text{if } a_{ij}^{t_p} \geq 0.5 \\ 0, & \text{otherwise} \end{array} \right\}$$

A soft thresholding power such as $t_p = 7$ (pow7) implies that genes with correlation less than approximately 0.906 will have their corresponding entry of B set to zero. Higher powers tend to result in a more sparse matrix of strongly correlated genes. By varying the t_p value for the same input data, we obtain graphs with different sizes and properties as illustrated in Table I. The analysis reported in this paper are based on $t_p = 7$.

WGCNA also requires specifying the graph network type: signed, unsigned, signed hybrid, or distance. We used the recommended type, i.e. signed hybrid (a hybrid of a weighted and unweighted network). The similarity is set to the correlation value if positive, otherwise it is equal to zero. The distance between two genes is inversely proportional to the correlation between their expression levels across all conditions. This means that two highly correlated gene will be geometrically close to each other, while uncorrelated genes will have a larger distance between them.

2) Construction of kNN Graphs: The kNN graphs were derived from the data using the Class Cover Catch Digraphs (CCCD) R package[1]. It creates a kNN graph such that an edge is added between vertices u and v if u is one of the k nearest neighbors to v, and v is one of the k nearest neighbors to u. Due to this symmetry requirement, each vertex will have a degree of at most k. Depending on the parameters used, kNN graphs tend to have fewer edges than geometric graphs created from the same data, as illustrated in Table I for $k = 5$ and $k = 30$.

The modularity measure computed for the various graphs is presented in Table I. Modularity quantifies the strength of *modules* (analogous to clusters) created when clustering a graph. A graph with high modularity has more than expected edges internal to its modules, and fewer than expected edges between modules. We applied modularity as a quick way to evaluate the "clusterability" of a graph. As can be observed from Table I, the kNN graphs had the higher modularity values with kNN-5 ranking the highest. However, from preliminary experiments, we observed that the lower k values resulted in graphs that were very sparse (not well-connected). Hence, we performed subsequent analysis on the kNN-30 graphs even though they were denser.

In previous works [1], [14] it was observed that, depending on parameters used, the kNN graphs were generally more efficient to cluster and yielded better results. This was probably due to the significantly smaller number of edges which made it easier to partition the graph with a smaller attack set. In this study, we apply clustering as a dimensionality space reduction tool (as described in Section III) and obtained better results with geometric graphs.

C. Grapevine Gene Expression Dataset

To identify differentially expressed genes in the two grapevine genotypes, *Vitis vinifera* (Cabernet sauvignon) and *Vitis aestivalis* (Norton), in response to the PM fungus, gene expression data [4], [6] was obtained by inoculating PM onto detached leaves of both Norton and Cabernet cultivars. One leaf was harvested from each plant and ten leaves were pooled as one sample at each time point. PM conidia-inoculated and mock-inoculated samples were ground separately in liquid nitrogen. Three sample replicates were processed for analysis. The gene expression data was generated on Affymetrix GeneChip *Vitis vinifera* genome array and extracted using the GeneChip operating software version 1.2. (The expression data is publicly available on the NCBI Gene Expression Omnibus website, accession number is GSE6404.)

The gene expression data utilized in this work consisted of 36 samples for each Norton and Cabernet cultivars i.e. 72 in total. The six different time points are initially (0 hour), and after 4, 8, 12, 24 and 48 hours. If any of the 72 samples for a particular feature (gene) was absent, that gene was removed

[1]Available at https://CRAN.R-project.org/package=cccd

from further consideration. A total of 9113 genes were 100% present or marginal across all conditions for both cultivars.

III. Methodology

The overall data-driven approach to identify features (genes) that significantly change in response to the PM inoculation as well as differ in pattern between Norton and Cabernet cultivars is illustrated in Figure 1. It consists of 5 key phases as described below.

A. Data Preprocessing

To investigate a suited data-driven ML framework, it is important to understand key characteristics of the problem at hand and how it translates to a computational intelligence framework. The preprocessing phase takes care of all necessary modifications that need to be addressed to translate the data effectively. Since we are interested in the differences or similarities in the expression patterns of Cabernet vs. Norton over time, we compute the mean differences in expression values between the inoculation and mock treatments at each of the six time-points. Prior to computing the mean differences, the three biological replicas at each time point were aggregated using geometric mean, given by

$$\bar{d} = \sqrt[n]{\prod_{i=1}^{n} d_i} \tag{2}$$

where d consists of all n replicas from a given timepoint, and d_i represents the data point from replica i. We also separated the input matrix into two separate data files since we were interested in comparing Cabernet against Norton. Thus, for subsequent analysis, we had two data matrices (Cabernet and Norton) with 6 conditions (representing each time point), a significantly reduced matrix from an input matrix of 72 conditions.

B. Graph Representation

During this phase, the preprocessed data is converted to a graph form. As mentioned in Section II-B, we explored two different graph representations: geometric and kNN graphs using the mean difference between inoculated and mock values, for both Norton and Cabernet. In this study, we selected the geometric graph with $t_p = 7$ (pow7) given that it was the least dense of the geometric graphs along with the kNN-30 graphs. The denser the graph, the greater the computational complexity.

C. Cluster Analysis

We conducted cluster analysis on the graphs using the NBR-Clust clustering framework with integrity resilience measure, as described in section II-A. The cluster analysis returns sub-partitions (clusters) of graph that represent genes that follow similar patterns of behavior i.e., the genes react similarly in terms of the mean difference between the inoculation and mock values across the six different time points (conditions). The question then is: how do we merge the clustering results obtained from the Cabernet graphs with the Norton graphs? How do we determine which are the biologically relevant clusters? We address these questions in the next phase.

D. Contrarian Approach

When clustering a very large dataset such as gene expression data, it is known that the relevant sets of genes are usually very small in comparison to the initial data size. Most genes behave similarly, in accordance with basic functions of the cell. Usually, the interesting subset of genes is much smaller. The contrarian approach, as discussed in [15], applies clustering as a tool to filter out a majority of non-essential (or non-interesting) genes and focuses on the outliers or the smaller clusters. It is assumed that the large clusters group the non-essential/basic functional genes together. As denoted by the term "contrarian approach" meaning "contrary", we are interested in the small clusters of genes that act "contrary" to the generally observed patterns. Our hypothesis is that these clusters will contain genes that do not follow general patterns, hence, could display distinct expression patterns of PM-induced transcripts across the time points that may insinuate biological relevance. We compared the smaller clusters obtained in Norton in contrast with the ones from Cabernet in terms of the genes that clustered in common between both (intersection of sets) as well as the differences of the sets.

E. Biological Analysis

To determine the biological relevance of the results obtained, we employed the Database for Annotation, Visualization and Integrated Discovery (DAVID) online tool [16], [17] as an initial method to observe the biological relevance of the results. Given a list of genes, the DAVID tool returns sets of genes that belong to known enriched functional-related gene groups and pathways among the input genes. Possible relevant pathways of genes that confer resistance to PM in Norton can include *catabolic process, biosynthesis, kinase, plant pathogen,* and *membrane*. We compared the means and standard deviations across time periods to determine if the differences in the clusters of patterns observed are significant. For future work, we plan to incorporate more rigorous statistical tests to compare differences in means across all time points between both species

IV. Results and Analysis

Cluster analysis on the geometric (pow7) graphs vs. the kNN-30 graphs yielded different results as shown in Table II. It is interesting to note that for both types of graphs, the size of the critical attack set is very large. Critical attack set nodes could consist of overlapping nodes and outliers. We plan to analyze the critical attack set in more details in future work. The geometric graphs produced one smaller size cluster (size 604) on the Norton dataset and 3 other large clusters (> 1700). The Cabernet graph seemed to partition readily into more clusters: about 10 with 6 of them less than 300 genes. The kNN graphs resulted in a higher number of clusters for both the Cabernet and Norton graphs with a close range of size distribution. The geometric graphs also seemed to result in

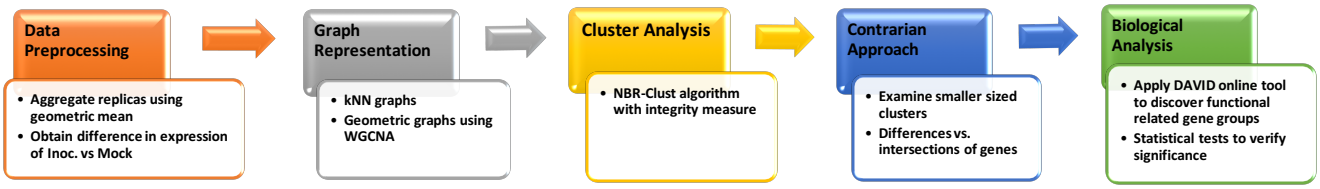


Fig. 1. Data-driven methodological framework to mine grapevine gene expression data using graph-based NBR-Clust algorithm.

TABLE II
CLUSTER DISTRIBUTION OBTAINED ON GEOMETRIC (POW7) GRAPHS COMPARED TO KNN-30 GRAPHS.

	kNN-30		**Geometric (pow7)**	
Cluster	Cabernet	Norton	Cabernet	Norton
0	430 (15)	524 (38)	282 (9)	604 (37)
1	622 (38)	748 (33)	2306 (138)	1966 (95)
2	806 (46)	707 (30)	347 (11)	2096 (85)
3	844 (32)	643 (17)	2297 (131)	1713 (132)
4	988 (56)	773 (60)	248 (11)	
5	428 (40)	472 (29)	273 (22)	
6	869 (56)	555 (20)	142 (8)	
7	218 (9)	382 (19)	68 (1)	
8	645 (32)	248 (14)	96 (5)	
9	77 (3)	397 (14)	21 (0)	
10		181 (17)		
11		62 (14)		
12		82 (7)		
S*	3174 (163)	3338 (179)	3030 (155)	2731 (142)

* S denotes the critical attack set nodes. Clusters with 3 or fewer genes are omitted for brevity.
Numbers in parenthesis indicate number of genes in the cluster that overlap with known429 genes.

more biologically relevant results as demonstrated in Tables III, IV and V.

Fung et al. identified a set of 625 genes that were differentially expressed between PM-inoculated and mock-inoculated in Cabernet [4]. Of these 625 genes, 491 were a subset of the 9113 genes explored in this work. In Table II, numbers in parenthesis indicate the number of genes in each cluster that overlap with the known subset of 491 genes. We would expect to find these 491 genes throughout all the clusters. They are genes that respond in some way to the presence of the disease, either they respond positively (up-regulated, higher expression) or negatively; and over time the response may vary. Both graphs had approximately the same percentage of those previously identified genes in the critical attack set across both cultivars. We will explore the biological relevance of the critical attack set in future analysis. Note that if a cluster does not have any of these 491 genes from prior work doesn't imply that it is not relevant. The relevance is confirmed by the biological analysis of the genes belonging to its cluster.

We computed set differences of the smaller sized clusters in both Cabernet and Norton graphs. The emphasis is on genes that are clustered together in Norton but not in Cabernet, given that the prior work had observed differences in the Cabernet (625 genes) compared to only a few (3 genes) in Norton. These genes were analyzed by the DAVID tool. The results revealed several genes that matched important keywords related to possible functional pathways that could influence plant disease resistance mechanism. Figures 2 and 3 illustrate a graph network visualization of the set difference between two interesting clusters, one of Norton and one of Cabernet. The important genes related to key pathways are denoted by different colors for ease of visualization. In Figures 2 and 3, the red nodes represent genes identified with *catabolic process*, orange nodes: *biosynthesis*, yellow nodes: *plant pathogen*, green nodes: *kinase*, and cyan nodes: *membrane* respectively. Nodes not matched with a pathway of interest are colored in gray.

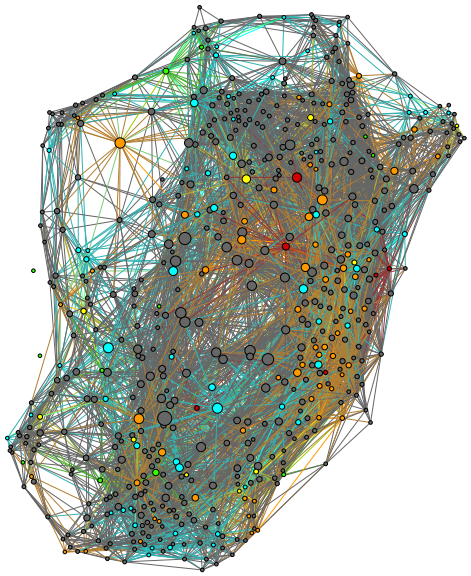


Fig. 2. Visualization of the set difference of Norton cluster 0 and Cabernet cluster 5 based on geometric graph. The different colors denote specific functional pathways; Orange nodes: biosynthesis, red nodes: catabolic process, cyan nodes: membrane, green nodes: kinase, and yellow nodes: pathogen.

Nodes in the figures are sized according to betweenness centrality, a measure of importance in a network. Genes with higher betweenness centrality are larger. It is interesting to note that the biosynthesis nodes tend to cluster together, tend to have higher betweenness centrality, and in some cases groups of biosynthesis nodes are matched with a corresponding membrane node. In contrast, many membrane nodes are distributed throughout the graph, and many have a low importance in the network, as expected. The catabolic process genes in red also tend to be important, but there are fewer of them, and the pattern is harder to distinguish. The corresponding network visualization results for the kNN-30 graphs, displayed in Figure 4, do not seem to yield any significant results unlike the geometric graphs.

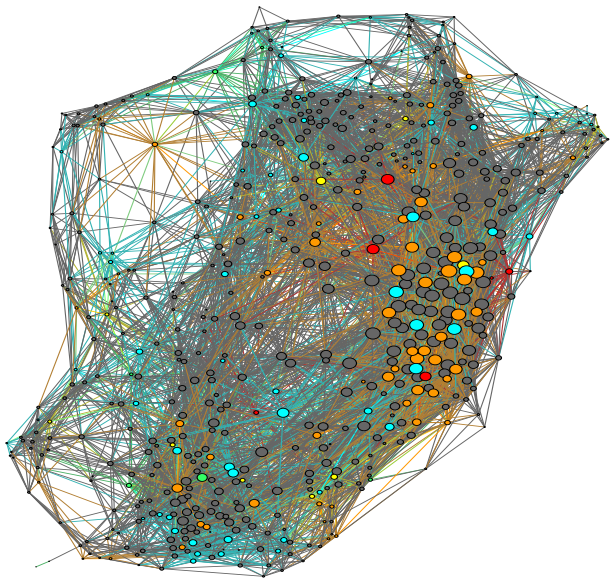


Fig. 3. Visualization of the set difference of Norton cluster 0 and Cabernet cluster 6 based on geometric graph. The different colors denote specific functional pathways; Orange nodes: biosynthesis, red nodes: catabolic process, cyan nodes: membrane, green nodes: kinase, and yellow nodes: pathogen.

The results from the DAVID tool comparisons for the geometric graphs are summarized for a few of the cluster comparisons in Tables III, IV, and V. The top row for each pathway denotes the mean and standard deviation values for Norton while the second row denotes the results for Cabernet. As can be observed, across all the tables, Norton generally appeared to have a lower variance around the mean. The difference in mean analysis for the geometric graph suggests significant differences in patterns observed across Norton in contrast to Cabernet. The results obtained by applying the DAVID tool on the kNN graphs (Tables VI and VII) did not yield interesting results. Figure IV illustrates the trends between Norton and Cabernet for certain genes belonging to the plant pathogen pathway found in cluster 0 of Norton (N0) but not in cluster 5 of Cabernet (C5). We observe varied differences in Cabernet at these genes, as they belong to multiple different clusters. The next steps will be to apply rigorous statistical tests to validate these differences.

V. Conclusion

This paper investigated a data-driven computational model, utilizing a graph-based clustering algorithm – Node-Based Resilience Clustering (NBR-Clust), to analyze grapevine gene expression data to identify possible gene biomarkers associated with PM disease defense mechanisms. We explored two graph representations (geometric and kNN) on the mean differences of PM inoculated vs. mock inoculated gene expression values of Cabernet and Norton (PM disease resistant) cultivars across 6 time points. The objective was to analyze the gene profiles obtained from mock inoculated vs. PM inoculated to identify sets of gene features responsible for the disease resistance mechanism of the Norton cultivar in contrast to Cabernet. By applying the contrarian approach, we hypothesized that smaller sized clusters will contain genes that do not follow general patterns, hence, could display distinct expression patterns of PM-induced transcripts across the time points that may insinuate biological relevance. Smaller clusters obtained in Norton were compared in contrast with the ones from Cabernet in terms of the genes that clustered together in one but not the other i.e. differences of the sets. The results obtained demonstrated the usefulness of the geometric graphs for this domain application in contrast to the kNN graphs. We also identified some genes that differ in patterns between Norton and Cabernet species that would be further explored to validate usefulness as potential biomarkers.

In future works, we will further analyze the biological significance of the genes discovered with this approach using more rigorous statistical analysis to characterize their impact in relation to the PM disease defense mechanisms.

References

[1] J. Matta, T. Obafemi-Ajayi, J. Borwey, D. Wunsch, and G. Ercal, "Robust graph-theoretic clustering approaches using node-based resilience measures," in *2016 IEEE 16th International Conference on Data Mining (ICDM)*, Dec 2016, pp. 320–329.

[2] M. Fasoli, S. Dal Santo, S. Zenoni, G. B. Tornielli, L. Farina, A. Zamboni, A. Porceddu, L. Venturini, M. Bicego, V. Murino *et al.*, "The grapevine expression atlas reveals a deep transcriptome shift driving the entire plant into a maturation program," *The Plant Cell Online*, vol. 24, no. 9, pp. 3489–3505, 2012.

[3] W. Qiu, A. Feechan, and I. Dry, "Current understanding of grapevine defense mechanisms against the biotrophic fungus (erysiphe necator), the causal agent of powdery mildew disease," *Horticulture research*, vol. 2, p. 15020, 2015.

[4] R. W. Fung, M. Gonzalo, C. Fekete, L. G. Kovacs, Y. He, E. Marsh, L. M. McIntyre, D. P. Schachtman, and W. Qiu, "Powdery mildew induces defense-oriented reprogramming of the transcriptome in a susceptible but not in a resistant grapevine," *Plant physiology*, vol. 146, no. 1, pp. 236–249, 2008.

[5] D. Pap, S. Riaz, I. B. Dry, A. Jermakow, A. C. Tenscher, D. Cantu, R. Oláh, and M. A. Walker, "Identification of two novel powdery mildew resistance loci, ren6 and ren7, from the wild chinese grape species vitis piasezkii," *BMC plant biology*, vol. 16, no. 1, p. 170, 2016.

[6] R. W. Fung, W. Qiu, Y. Su, D. P. Schachtman, K. Huppert, C. Fekete, and L. G. Kovács, "Gene expression variation in grapevine species vitis vinifera l. and vitis aestivalis michx." *Genetic Resources and Crop Evolution*, vol. 54, no. 7, pp. 1541–1553, 2007.

[7] K. C. Amrine, B. Blanco-Ulate, S. Riaz, D. Pap, L. Jones, R. Figueroa-Balderas, M. A. Walker, and D. Cantu, "Comparative transcriptomics of central asian vitis vinifera accessions reveals distinct defense strategies against powdery mildew," *Horticulture Research*, vol. 2, p. 15037, 2015.

[8] J. Matta, G. Ercal, and J. Borwey, "The vertex attack tolerance of complex networks," *RAIRO-Operations Research*, vol. 51, no. 4, pp. 1055–1076, 2017.

[9] C. Barefoot, R. Entringer, and H. Swart, "Vulnerability in graphs-a comparative survey," *Journal of Combinatorial Mathematics and Combinatorial Computing*, vol. 1, pp. 12–22, 1987.

[10] P. G. Drange, M. S. Dregi, and P. vant Hof, "On the computational complexity of vertex integrity and component order connectivity," in *Algorithms and Computation*. Springer International Publishing, 2014, pp. 285–297.

[11] P. Langfelder and S. Horvath, "Wgcna: an r package for weighted correlation network analysis," *BMC bioinformatics*, vol. 9, no. 1, p. 559, 2008.

[12] ——, "Fast r functions for robust correlations and hierarchical clustering," *Journal of statistical software*, vol. 46, no. 11, 2012.

[13] L. Song, P. Langfelder, and S. Horvath, "Comparison of co-expression measures: mutual information, correlation, and model based indices," *BMC bioinformatics*, vol. 13, no. 1, p. 328, 2012.

TABLE III

MEAN AND STANDARD DEVIATION COMPARISONS FOR THE SET DIFFERENCES OF GEOMETRIC GRAPH (NORTON 0 (UPPER ROW) VS CABERNET 5 (LOWER ROW)) ACROSS CERTAIN BIOLOGICAL PATHWAYS.

Pathway	0 hours	4 hours	8 hours	12 hours	24 hours	48 hours
biosynthesis	13.5 ± 66.91	-3.15 ± 61.07	35.35 ± 71.84	-80.54 ± 93.07	-133.98 ± 134.67	172.57 ± 195.3
	-129.46 ± 269.39	-255.63 ± 320.23	-16.91 ± 161.05	-40.85 ± 154.05	-229.99 ± 296.33	-313.07 ± 264.84
catabolic process	46.65 ± 32.78	40.88 ± 55.35	115.6 ± 115.54	-71.69 ± 59.97	-87.56 ± 97.1	175.38 ± 141.31
	58.57 ± 73.43	-80.76 ± 93.51	108.68 ± 99.05	36.13 ± 103.82	95.52 ± 92.15	25.5 ± 130.82
kinase	30.0 ± 66.87	0.43 ± 25.46	17.93 ± 20.89	4.41 ± 29.12	-15.29 ± 41.19	11.81 ± 37.97
	50.14 ± 72.32	2.78 ± 45.41	9.25 ± 21.54	38.83 ± 71.52	4.34 ± 45.28	-5.72 ± 71.46
membrane	0.97 ± 55.58	-16.93 ± 57.28	19.09 ± 56.36	-38.32 ± 81.2	-72.06 ± 131.5	104.36 ± 194.74
	-50.0 ± 222.43	-139.41 ± 286.29	-11.13 ± 126.29	-21.05 ± 116.5	-86.33 ± 230.9	-135.47 ± 252.24
pathogen	29.39 ± 113.46	-33.01 ± 79.04	29.52 ± 81.55	-43.78 ± 73.96	-116.39 ± 144.39	163.12 ± 170.11
	1.06 ± 226.97	-199.1 ± 324.15	25.54 ± 128.89	-21.62 ± 159.01	-170.32 ± 287.52	-169.47 ± 284.2

TABLE IV

MEAN AND STANDARD DEVIATION COMPARISONS FOR THE SET DIFFERENCES OF GEOMETRIC GRAPH (NORTON 0 (UPPER ROW) VS CABERNET 6 (LOWER ROW)) ACROSS CERTAIN BIOLOGICAL PATHWAYS.

Pathway	0 hours	4 hours	8 hours	12 hours	24 hours	48 hours
biosynthesis	10.92 ± 69.1	-4.49 ± 61.14	36.32 ± 72.12	-80.42 ± 92.27	-138.84 ± 136.35	174.41 ± 194.17
	-131.46 ± 268.01	-262.97 ± 320.75	-23.39 ± 164.58	-44.05 ± 153.93	-239.24 ± 297.96	-315.31 ± 261.59
catabolic process	46.65 ± 32.78	40.88 ± 55.35	115.6 ± 115.54	-71.69 ± 59.97	-87.56 ± 97.1	175.38 ± 141.31
	58.57 ± 73.43	-80.76 ± 93.51	108.68 ± 99.05	36.13 ± 103.82	95.52 ± 92.15	25.5 ± 130.82
kinase	24.14 ± 36.12	4.24 ± 13.36	13.01 ± 9.58	-6.62 ± 12.15	-9.09 ± 23.16	6.54 ± 36.6
	20.6 ± 47.41	-7.22 ± 49.94	4.17 ± 13.45	3.04 ± 15.1	-8.73 ± 43.16	-34.45 ± 57.62
membrane	-4.6 ± 50.74	-14.28 ± 53.27	22.62 ± 51.92	-44.8 ± 78.34	-76.07 ± 131.59	100.08 ± 191.04
	-57.42 ± 216.64	-141.78 ± 285.07	-8.3 ± 115.65	-30.42 ± 115.28	-95.51 ± 226.32	-135.89 ± 253.03
pathogen	29.39 ± 113.46	-33.01 ± 79.04	29.52 ± 81.55	-43.78 ± 73.96	-116.39 ± 144.39	163.12 ± 170.11
	1.06 ± 226.97	-199.1 ± 324.15	25.54 ± 128.89	-21.62 ± 159.01	-170.32 ± 287.52	-169.47 ± 284.2

TABLE V

MEAN AND STANDARD DEVIATION COMPARISONS FOR THE SET DIFFERENCES OF GEOMETRIC GRAPH (NORTON 0 (UPPER ROW) VS CABERNET 8 (LOWER ROW)) ACROSS CERTAIN BIOLOGICAL PATHWAYS.

Pathway	0 hours	4 hours	8 hours	12 hours	24 hours	48 hours
biosynthesis	10.45 ± 68.68	-4.39 ± 60.67	35.83 ± 71.68	-79.43 ± 91.92	-137.14 ± 136.01	171.98 ± 193.69
	-129.77 ± 266.32	-258.97 ± 319.94	-23.22 ± 163.33	-43.69 ± 152.79	-236.06 ± 296.81	-310.38 ± 262.62
catabolic process	46.65 ± 32.78	40.88 ± 55.35	115.6 ± 115.54	-71.69 ± 59.97	-87.56 ± 97.1	175.38 ± 141.31
	58.57 ± 73.43	-80.76 ± 93.51	108.68 ± 99.05	36.13 ± 103.82	95.52 ± 92.15	25.5 ± 130.82
kinase	24.14 ± 36.12	4.24 ± 13.36	13.01 ± 9.58	-6.62 ± 12.15	-9.09 ± 23.16	6.54 ± 36.6
	20.6 ± 47.41	-7.22 ± 49.94	4.17 ± 13.45	3.04 ± 15.1	-8.73 ± 43.16	-34.45 ± 57.62
membrane	-5.55 ± 53.33	-14.98 ± 55.95	27.14 ± 53.05	-50.09 ± 81.07	-84.73 ± 137.11	115.98 ± 197.07
	-63.5 ± 227.77	-159.97 ± 297.1	-7.06 ± 121.41	-33.31 ± 120.71	-101.8 ± 236.42	-157.35 ± 260.84
pathogen	29.39 ± 113.46	-33.01 ± 79.04	29.52 ± 81.55	-43.78 ± 73.96	-116.39 ± 144.39	163.12 ± 170.11
	1.06 ± 226.97	-199.1 ± 324.15	25.54 ± 128.89	-21.62 ± 159.01	-170.32 ± 287.52	-169.47 ± 284.2

TABLE VI

MEAN AND STANDARD DEVIATION COMPARISONS FOR THE SET DIFFERENCES OF KNN GRAPH (NORTON 10 (UPPER ROW) VS CABERNET 10 (LOWER ROW)) ACROSS CERTAIN BIOLOGICAL PATHWAYS.

Pathway	0 hours	4 hours	8 hours	12 hours	24 hours	48 hours
biosynthesis	10.34 ± 46.37	-7.8 ± 38.88	4.11 ± 26.19	11.46 ± 47.16	-29.29 ± 42.93	-0.23 ± 50.95
	10.55 ± 110.13	-7.73 ± 63.08	2.91 ± 58.38	-34.43 ± 75.44	-45.48 ± 115.65	-28.41 ± 77.99
kinase	21.75 ± 35.01	-22.48 ± 18.21	3.27 ± 10.72	17.65 ± 23.84	-10.13 ± 8.42	-6.8 ± 11.48
	11.48 ± 60.53	-14.26 ± 27.92	-16.02 ± 48.57	25.66 ± 81.45	24.96 ± 40.1	4.48 ± 31.5
membrane	37.16 ± 106.37	-15.29 ± 21.79	1.46 ± 16.46	10.69 ± 28.49	-22.68 ± 45.06	-17.09 ± 34.6
	45.02 ± 109.81	4.77 ± 35.87	3.86 ± 30.72	-20.53 ± 44.97	-21.45 ± 63.42	-5.2 ± 29.99

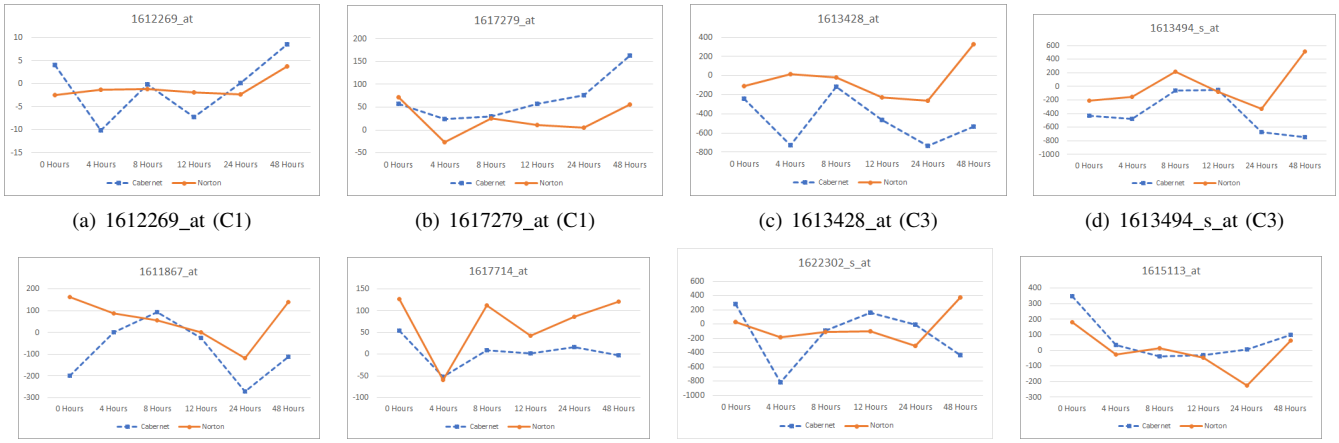
TABLE VII

MEAN AND STANDARD DEVIATION COMPARISONS FOR THE SET DIFFERENCES OF KNN GRAPH (NORTON 12 (UPPER ROW) VS CABERNET 9 (LOWER ROW)) ACROSS CERTAIN BIOLOGICAL PATHWAYS.

Pathway	0 hours	4 hours	8 hours	12 hours	24 hours	48 hours
biosynthesis	-9.17 ± 13.82	15.45 ± 35.73	-5.66 ± 17.28	-1.34 ± 32.35	-17.01 ± 36.96	6.9 ± 46.6
	-27.24 ± 46.63	-1.51 ± 41.1	-18.48 ± 40.58	-34.53 ± 56.53	-55.5 ± 118.71	5.32 ± 25.71
kinase	0.24 ± 16.58	18.6 ± 20.13	1.72 ± 9.96	-16.98 ± 19.0	16.85 ± 7.75	5.04 ± 19.44
	-8.92 ± 17.35	0.24 ± 34.32	2.84 ± 16.43	25.19 ± 25.9	7.39 ± 33.94	13.49 ± 42.22
membrane	-3.65 ± 10.95	4.0 ± 13.31	-5.05 ± 11.96	-8.92 ± 14.93	5.63 ± 21.16	-4.57 ± 18.1
	-9.39 ± 18.13	5.97 ± 14.93	0.08 ± 11.58	0.14 ± 24.47	8.65 ± 46.0	21.33 ± 56.84

(a) Norton 10 vs. Cabernet 9 (b) Norton 11 vs. Cabernet 9 (c) Norton 12 vs. Cabernet 9

Fig. 4. Three cluster comparisons of set differences between Norton and Cabernet based on the kNN-30 graph.



(a) 1612269_at (C1) (b) 1617279_at (C1) (c) 1613428_at (C3) (d) 1613494_s_at (C3)

(e) 1611867_at (C-S) (f) 1617714_at (C-S) (g) 1622302_s_at (C4) (h) 1615113_at (C2)

Fig. 5. Trend in difference of expression levels across the 6 time points (0, 4, 8, 12, 24, 48 hrs) for selected genes belonging to Norton cluster 0 (N0) but not in Cabernet cluster 5 (C5). The blue dashed lines represent Norton while the orange solid lines - Cabernet. Below each chart, we specify the cluster each gene is found in the corresponding Cabernet graph. C-S denotes critical attack set of Cabernet.

[14] J. Borwey, D. Ahlert, T. Obafemi-Ajayi, and G. Ercal, "A graph-theoretic clustering methodology based on vertex-attack tolerance," in *The Twenty-Eighth International Flairs Conference*, 2015.

[15] T. T. Toma, Z. Williams, J. Dawson, and D. Adjeroh, "What can one chromosome tell us about human biogeographical ancestry?" in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2017.

[16] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using david bioinformatics resources," *Nature protocols*, vol. 4, no. 1, p. 44, 2008.

[17] ——, "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists," *Nucleic acids research*, vol. 37, no. 1, pp. 1–13, 2008.