petal: A Novel Co-Expression Network Modeling System

Juli Petereit, Frederick C. Harris, Jr., and Karen Schlauch
University of Nevada, Reno
Reno, NV, USA
julipetereit@gmail.com Fred.Harris@cse.unr.edu schlauch@unr.edu

Abstract— **With the introduction of microarray technology to measure gene expression in the late 1990's and the advent of current next-generation whole-genome and whole-transcriptome sequencing technology, fast and robust tools are needed to examine, identify, and study relations between genes and proteins at the systems level. Networks provide effective models to study complex systems, including complex biological systems, such as gene and protein interaction networks. We introduce a novel approach to generate gene co-expression network models based on experimental gene expression measures. This approach includes statistical, mathematical, and biological considerations not highlighted by existing co-expression analysis tools. First, most high-throughput expression data are not normally distributed, yet many available network approaches are based on parametric methods. Here appropriate metrics are provided to generate statistically sound network models. Secondly, most biological networks are known to have approximate scale-free and small-world structure, and the biological networks built here follow both these two properties. Thirdly, to generate these small-world, scale-free network models, user-selected input parameters are not required, thereby leading to reproducible results. Lastly, this approach is designed for high-throughput whole-systems data. This method is implemented in the programming language R. Its application to several whole-genome experimental datasets has generated novel meaningful results useful for further biological investigation.**

Keywords—small-world, scale-free, R, whole-omics approach, parameter-free algorithm

I. Introduction

Within the life/biomedical sciences, high-throughput technologies produce large experimental omics datasets at overwhelming speeds. Analysts are left to organize, structure, and analyze these data in sufficient and efficient ways. Computational biology, bioinformatics, systems biology, network biology, and network medicine offer interdisciplinary tools to help solve these challenges. Here, our focus is the efficient analysis of high-throughput gene expression data from microarrays and next-generation sequencing platforms (RNA-seq) via network approaches.

Applications of networks and their analysis have become standard tools in the systems biology toolbox for their versatility and powerful approach to whole-system analysis [1, 2, 3]. Networks provide an effective approach to analyze very large complex datasets. The biological networks of interest here are co-expression networks. A co-expression network is built from gene expression data collected over a series of experimental conditions. Vertices (nodes) correspond to genes, and edges represent a pre-defined relationship between them. With a network at hand, its topological properties aid in revealing whole-system expression patterns, putative gene interactions, potential functional groupings, the association of functions to genes of unknown function, and possible regulations within the system. Examining network properties in combination with well-defined testing hypotheses can lead to the identification of putative key players within a pathway and thus possible drug targets in future research.

This paper introduces a novel network analysis tool called **petal**. First we outline the typical approaches to co-expression network analysis and discuss a few areas that may be strengthened by the use of **petal**. Then the main elements of the tool's algorithm are presented, followed by a demonstration of **petal** to analyze a whole-genome experiment. We show that **petal** is a user-friendly and laborsaving approach, providing biologically meaningful outcomes and easier user-accessibility to tangible results than many other available co-expression methods.

II. Standard Co-Expression Network Approach

A. Co-expression Networks

Co-expression networks are based on experimental expression measures of m gene identifiers across n conditions (treatments/time points/replicates). Each gene is represented by a vertex. Genes are connected by an edge if their expression measures across the n conditions are similar to a certain degree. Figure 1a shows a simple example of a small network graph, and Figure 1c demonstrates a group of genes with similar expression across 28 measures. Mathematically, the expression profile of a gene is an n-dimensional vector; association between two vectors is defined with a metric and a threshold. Association between each gene pair is computed via the metric, transforming the $m \times n$ expression matrix into an $m \times m$ symmetric similarity matrix, representing a completely connected graph with vertices connected at different strengths. Next, an adjacency function transforms all gene pair association measures into a resulting unweighted or weighted network, which is mathematically presented by the adjacency (incidence) matrix. The adjacency matrix of an unweighted network is binary; the measure of each gene pair is in a "connected: 1" or "not-connected: 0" state defined by a user-specified threshold imposed on the association metric. In a

weighted network, all vertices stay connected at different weights which are calculated via an adjacency function.

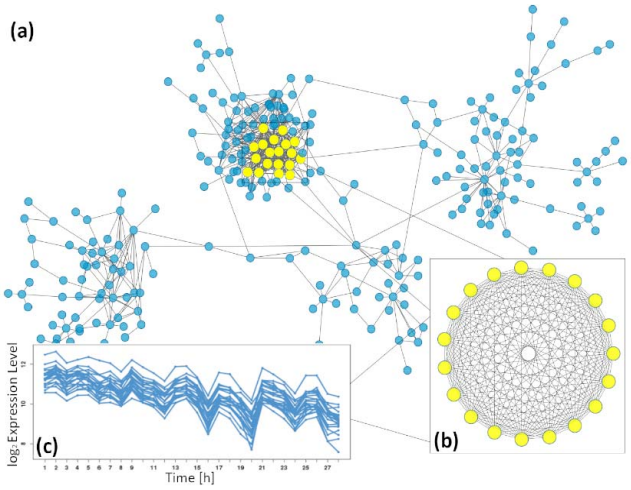

Fig. 1: Small Network Analysis Example (a) Network with small-world and scale-free architecture and highlighted 19-vertex clique (b) extracted clique (c) gene expression profiles of the 19 clique members.

The underlying assumption of co-expression network analysis is that genes with similar expression patterns are possibly co-expressed, co-regulated, share common functionality, and/or might be regulated by a joint transcription factor (TF). Consequently, groups of similar expression profiles across experimental conditions can be hypothesized to share common functionality by means of the 'Guilt-by-Association' principle [4].

B. Edge Definition: Metric and Threshold

Association between expression vectors is determined by a metric such as a measure of correlation or geometric distance and a user-defined threshold. A commonly used metric is the Pearson Correlation Coefficient (PCC) with a threshold of 0.8 [5, 6]. The PCC is a parametric measure, and should only be applied to normally distributed (expression) data. A robust alternative to the PCC is the Spearman Correlation Coefficient (SP), which is used in multiple studies [7, 8]. Distance metrics can also be utilized, e.g., Euclidean- and Manhattan distances; these non-parametric measures do not provide insight into the behavioral changes of expression profiles and are based solely on magnitude. Mutual Information (MI) is another non-parametric metric based on entropy and can process missing data values better than other metrics [9, 10]. Association can be measured by a number of other, less common metrics; for more detail on different measures refer to [11, 12, 13, 14].

Pairwise association measures are transformed into an adjacency matrix according to a user-specified adjacency function and threshold. The most simple adjacency function is a discrete transformation that converts the expression association measures to 1 or 0 depending upon a user-selected threshold, to indicate similar expression or not, respectively. This transformation is called the Signum Adjacency Function [15] and is defined in (1), where the variable a represents the association measures and τ is the threshold on which to define association, i.e., an edge between vertices. Note that by definition in (1), the association measure a is a similarity metric, with highest possible numeric value indicating the strongest association. When a is a distance measure, the inequality signs in (1) are reversed.

$$signum(a) = \begin{cases} 0 & if\ a < \tau \\ 1 & if\ a \geq \tau \end{cases} \quad (1)$$

C. Network Topology

Topological properties of a network are robust descriptive measures that objectively describe the network's architecture. Such measures include vertex degree distribution, cluster coefficient, path length, diameter, density, and others [16]. The degree of a vertex is the number of its neighbors. A vertex's cluster coefficient indicates how well its neighbors are connected: when a vertex has a cluster coefficient of 1 then all of its neighbors are connected to each other. The path length between two vertices is the number of edges within their shortest path. The diameter of a network is the shortest path between the vertices that are furthest apart (shortest longest path) within the network. Density is the fraction of the number of existing edges in the network by the number of total edges possible. Common architectural features found in complex biological networks, including co-expression networks, are small-world and scale-free characteristics [17, 18]. These two properties have been proven to be standard characteristics of true complex biological networks [19, 20, 21].

1) Scale-free

Albert-László Barabási and Réka Albert inaugurated the notion of a scale-free network in 1999, and showed that most complex systems including biological complex systems are realistically modeled by networks following this property [17]. In a scale-free network, there are many vertices with few connections and only few vertices with a large number of connections. A network is defined to have scale-free architecture when the degree distribution of the vertices follows a power-law distribution [16, 22, 23].

2) Small-world

For a network model to be small-world it must be made of densely connected subnetworks that are linked together in such a fashion that the path between any vertex pair is relatively short [18]. To categorize a network as small-world, its average path length needs to be relatively short in comparison to random network models or by chance alone. This phenomenon is often referred to as "six-degrees of separation" [18, 24]. Secondly, a small-world network's cluster coefficient must be larger than in a random graph.

3) Network Components

A network component is a set of vertices that are connected by paths. If a network is made of one component it is considered a connected network. If a network model has two components, then this model has two disjoint subnetworks and not every vertex has a path to every other vertex within the entire network model. Network architectures, scale-free and small-world, are defined under the assumption that the network is connected; however, their defining topological properties (vertex degree distribution, average cluster coefficient, average

path length) can be calculated without this assumption by excluding vertex pairs in different components when calculating averages. It is seldom the case for biological network models based on expression data to be one single component. The biggest component of a multi-component network must include at least 90-98% of the network's vertices for the topological measures to reliably define the model's architecture; otherwise the topological measures can lead to misinterpretation [16].

D. Structures within Networks

One of the goals of co-expression network analysis is to extract structures (subnetworks, paths) from the entire network and examine these for biological patterns or association.

1) Modules, Cluster, and Communities

Clusters, modules, and communities are loosely defined terms and are interchangeably used in the literature [16, 25, 26]. They describe subnetworks that are (relatively) tightly connected. Often hierarchical clustering is performed on the pairwise-distance matrix to organize the networks into hierarchical trees which then can be cut at a user-specified height to obtain network modules. Module detection is the general practice in defining (tightly) connected gene groups or partitioning the network into smaller subnetworks [14, 16, 25, 26].

2) Cliques

Cliques are completely connected subnetworks, and share the same topological properties regardless of dimension: the diameter and cluster coefficient of a clique is always equal to one, and every vertex of the clique has the same degree. The members of a clique form an equivalence class which is graphically represented in Figure 1b. Genes in a clique follow the transitive property which results in less variation across clique members' expression profiles (Figure 1c) compared to groupings obtained from standard clustering routines [26]. The mathematical definition of a clique is: A subnetwork of j vertices is a clique if and only if the subnetwork has $j(j-1)/2$ number of edges. Extracting cliques from a network is a common network analysis step, but computationally very expensive and an NP-complete problem. The extraction of fully connected subgraphs is considered too stringent for some biological testing hypotheses and very time-consuming when the network is densely connected. To provide balance, a more relaxed version of cliques can be used, referred to as fuzzy cliques [14].

3) Vicinity Network

A vicinity network (VN) is a subnetwork representing the intermediate neighborhood of a single vertex or a completely connected group of vertices (clique). The VN of a clique includes all vertices to which every member in the clique is connected. The topological properties of VNs can vary greatly, but their extraction from a network is very fast. These smaller subnetworks can be examined more closely and cliques are extracted at a much smaller computational cost from VNs than from the entire network. Often, some precision is lost when computational time is decreased, there is no loss of information when gene-specific cliques are extracted from its vicinity network than when they are mined from the entire network.

E. Challenges of Co-Expression Approaches

The life scientist may experience several challenges while using co-expression networks methods. These can include, but are not limited to 1) the choice of a proper association metric relevant to the data distribution and experimental hypothesis; 2) the absence of explicit confirmation that the constructed network follows the scale-free and small-world properties; 3) the inconvenience of having to enter a large number of user-specified input variables; 4) the restriction of using only datasets attached to a tool's integrated database; 5) the distress of needing to use command-line programming to run the program; 6) the possibility that the program must be downloaded onto the user's desktop, thereby limiting the analysis to the power of the personal workstation.

PCC is the most common default metric in co-expression network tools. It is a convenient choice because scientists are familiar with it, and its computational cost is very low in comparison to MI, for example. The PCC is based on normality assumptions, and thus is not the proper choice for many datasets, such as RNA-seq data, which typically follow a negative binomial distribution [27]. Furthermore, correlation based association should not be confused with causation, or used to conclude that correlation necessarily implies co-regulation.

Co-expression networks are shown to have small-world and scale-free properties and can be realistically modeled by these two model structures [19, 20, 21], but to our knowledge there is no co-expression network method inheritably constructing networks with both these biological properties. For example, the Weighted Gene Co-expression Network Analysis (WGCNA) approach includes a scale-free topological fitting index to allow user-intervention in constructing a scale-free network, but WGCNA does not purposefully construct networks following small-world architecture [15, 26].

Many tools require user-specified parameters, such as the similarity metric threshold that defines edges. This threshold is fundamental, it is used in the governing steps of network construction and influences the structure of the model; results obtained are almost completely dependent on this threshold. Therefore, the threshold should be objectively set, rather than subjectively chosen by the user. Common practice is to define similarity with a correlation value of 0.8 and higher [5, 6]. However, there is no consensus on metric threshold values; it is more of an arbitrary selection that does not necessarily reflect biological relevance.

Other network analysis tools are associated to particular databases and can only generate a network from its data [28, 29]. Command-line tools are often uncomfortable or inconvenient for the basic science researcher. Downloading software onto a personal computer may be impeded by institutional security regulations, or the user's personal computer does not offer enough memory to complete the computations necessary to generate the network.

Finally, networks are generated in order to study experimental hypotheses, thus resulting structures containing genes of interest should be extracted and presented to the user in a clear and understandable manner. These results should be

both self-contained and easily transferable to a downstream analysis tool, such as Cytoscape [30] or Pajek [31].

III. THE **PETAL** APPROACH

The **petal** tool was designed to strengthen the standard flow of co-expression analysis. Upon the evaluation of current co-expression network tools [26, 28, 29, 32] our first goal was to develop a network construction algorithm that confirms the resulting model follows real biological network characteristics: scale-free and small-world. An additional goal was to generate network models based on entire omics expression datasets to ensure a true whole-transcriptome or whole-genome representation rather than a model based on a pre-selected subset of genes (e.g., differentially expressed (DE) genes). With no other user input but the experimental dataset, the construction of the network model is completely automated. The tool is implemented in the programming language **R** [33]. To further ease the utility of this application to the life scientist we propose to develop a graphical user interface (GUI) around the **R** program to invite the novice researcher to enjoy this network approach.

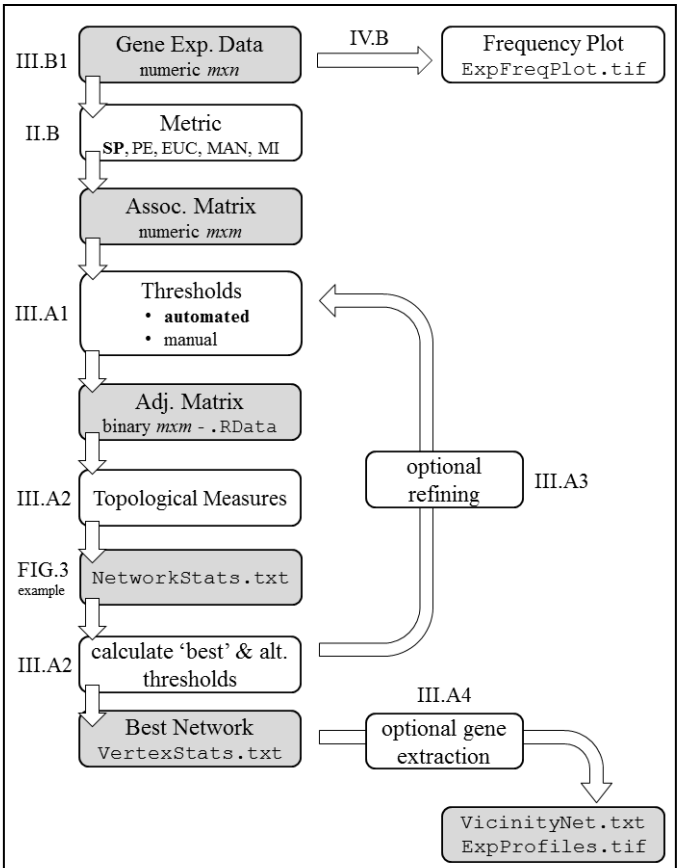

Fig. 2: Illustration of computational pipeline implemented in R Roman numbers refer to the section in the text in which the step is explained in detail; grey colored rectangles indicate data; white rectangles specify code; output files are writing in Courier New; Abbreviations: SP – Spearman Correlation, PE – Pearson Correlation, EUC – Euclidean Distance, MAN – Manhattan Distance, MI – Mutual Information, alt. – alternative.

A. Algorithm

The novelty of **petal** lies in its automated construction of scale-free and small-world network models. A summary of the computational pipeline is shown in Figure 2 that contains the main steps with references to the section in the text for more details.

1) Selection of threshold list

After the calculation of all pair-wise association measures, the $m(m-1)/2$ association measures are sorted from strongest to weakest association. For example, correlation and MI are organized in descending order, whereas distance measures are sorted in increasing order. For a network of m vertices to be connected, it must have at least $m-1$ edges, hence the first threshold, which is the most stringent, is set to the value at the $m\text{-}1^{st}$ position in the sorted association measure list. The last threshold is based on several empirical evaluations: In a series of actual RNA-seq and microarray whole-omics test datasets, network models with edges more than 150 times the number of vertices prove to be too dense for evaluation within reasonable computational runtime. Furthermore, in all observed cases of network models with this many edges, their corresponding vertex degree distributions do not follow a power-law function; consequently, the models cannot be classified as scale-free. To keep the runtime relatively low, we impose a restriction on all considered thresholds by limiting the number of edges to 150 times the number of vertices in the network. Consequently the last threshold is set to the value at the $150m^{th}$ position in the sorted association measure list. The interval between first and last threshold is split into six equal subintervals, resulting in the list of seven considered thresholds. The number of considered association thresholds can be increased to refine the length of subintervals, but this could lead to an exponential increase in runtime to construct the adjacency matrices for each threshold. Instead an optional refine step is integrated, which is described in *III.A.3*.

2) Selection of 'best' threshold

For each of these thresholds the adjacency matrix is generated, each of which corresponds to a network model. Topological measures relevant for defining small-world and scale-free are calculated for each model. Functions from the **R**-library 'igraph' [34] are utilized to obtain the average cluster coefficient, average path length, diameter, and the number of components. The average cluster-coefficient and average path length are used to determine small-worldness. To assure the network is scale-free, the networks' vertex degree distribution must follow a power-law distribution p_k, where k is the degree and C and α are positive constants. The power-law function is shown in (2). For each gene the number of its neighbors, k, is calculated to obtain the actual vertex degree distribution. The logarithmic transformation of the power-law function in terms of $\log(k)$ is linear as demonstrated in (3). To evaluate how well the power-law function (2) fits the actual degree distribution, the degree distribution is log-transformed.

$$p_k = Ck^{-\alpha} \tag{2}$$

$$\log(p_k) = -\alpha \log(k) + c \tag{3}$$

Linear regression is applied to the log-transformed degree distribution to measure how well the data follow the power-law distribution. The coefficient of determination (R^2) and the slope of the linear regression are computed and recorded. The slope α of the linear regression corresponds to the power in (2) and should lie within the interval (-3,-1) in order for the network to be scale-free [16, 23]. Topological properties obtained from the graph are used to identify the largest network component and its relative size. Lastly, vertices which are not connected to any other vertex are removed from consideration as they do not provide any information in terms of association. The percentage of remaining vertices is recorded. The resulting network models are weighted against each other based on their topological properties. The 'best' threshold is considered to have constructed a network model that is scale-free, small-world, with its biggest component including at least 95% of the network's vertices to confirm that most vertices are within one component, and retains the maximum number of vertices from the original dataset. If such a network cannot be identified, the user is alerted, and all network models with their topological measures remain accessible.

3) Refining threshold

Depending on the identified first and last thresholds, the interval between these two values can be relatively large. Consequently, the step sizes between considered thresholds are large and a 'better' threshold might be missed between the measured thresholds. To account for a large step size between threshold values, a refining step is included in the algorithm. Refining thresholds is an optional step, as this comes at a cost of longer runtime.

After the first round of initial threshold setting and identification, the 'best' threshold is not reported; instead, it is reused for a second round to test for scale-free and small-world. Let the 'best' threshold be denoted as t_{best}. To calculate a new list of thresholds with smaller step size, new first and last thresholds are needed. We differentiate between two cases. Case 1: besides t_{best} one or more thresholds meet the criteria of the algorithm, denoted as t_{alt}. Case 2: only t_{best} produces a scale-free, small-world network model. For Case 1 the possible alternative thresholds t_{alt} and t_{best} are sorted, the strongest and weakest associations are set to the new first and last thresholds. For the second case t_{best-1} and t_{best+1} are set to the first and last threshold, respectively. The new first and last thresholds cover a smaller spectrum, so the step size will be smaller and the choice of final threshold more precise. After the first and last thresholds are assigned, the interval between the two is again split into six equal subintervals, resulting in the list of refined thresholds. The algorithm then proceeds with *III.A.2*.

4) Extracting groups based on genes of interest

If a list of gene identifiers is provided by the user, then the algorithm continues after *III.A.2* by extracting vicinity networks from the identified 'best' network model. A (one-neighbor) vicinity network (VN) of vertex i is a subnetwork including vertex i and all its direct neighbors and their edges. There are two VN extraction options the user can select: 1) considering genes individually, or 2) combining the genes that are connected in the network model.

Option 1: The gene list is acknowledged and each gene and its neighbors are writing to file with the VN's density.

Option 2: The genes are first extracted from the network, and tested for connections by finding all maximal cliques within this extracted subnetwork. Specifically, for k genes of interest, the resulting network has dimension k by k. This subnetwork is investigated for all maximal cliques. Each maximal clique is treated separately while identifying its neighbors. Let there be s members in clique r, then clique r's neighbors are the common neighbors of its s members. Neighbors of each maximal clique are written to a file distinguishing between neighbors and the clique genes obtained from the user's identifiers. The density of the VN is also reported.

5) User Input

The user supplies the expression data file to **petal**. In addition, there are three optional steps: the selection of an association metric, user-specified thresholds (for the advanced user), and the upload of a list of genes which are of particular interest to the researcher (see the example in *IV*). To ensure that a statistically appropriate similarity metric is used to construct the network, **petal** provides the user with an estimate of their data distribution. Thus the user is able to identify whether data are approximately normally distributed. In this uncommon case, the user can then select a parametric similarity measure, such as the PCC. Other non-parametric association metrics currently implemented in **petal** include the Spearman Correlation Coefficient (SP), Euclidean and Manhattan Distance, and Mutual Information (MI). The SP is currently set as the default metric. The second optional step is to select up to five association thresholds instead of using the automated threshold computation. A network graph with thousands of vertices cannot illuminate the behavior of a small subgroup of genes of interest. As the third optional step, **petal** allows the user to examine the network structures of a few genes by extracting the one-neighbor vicinity networks of one or more genes independently or together.

6) User Output

Upon completion, **petal**'s accessible files include: general information file (.txt), network file (.txt), adjacency matrices (.RData), two topology tables (.txt), vicinity network files (.txt), and the expression profiles (.tif) of each vicinity network. The network file can be uploaded into Cytoscape directly. Cytoscape, an Open Source tool, can be used for visualization and offers several network viewing tools via various plugins [30, 35]. The .RData files of the network adjacency matrices are provided for convenient loading into **R**, enabling the advanced user to personalize downstream analysis if desired. In addition, the user can look at the characteristics of networks generated on different thresholds. Additionally, a table is provided which includes all network vertices with their degree and cluster coefficient. Each identified vicinity network is reported with gene membership, its density, and the number of edges missing to be a clique.

7) The **petal** *GUI*

For the novice user, we have designed a web-based graphical user interface (GUI). This allows the researcher to input data and optional parameters via a simple interface

without requiring any packages or software to be downloaded. Currently the **petal** GUI is undergoing alpha testing within our institution. (Please see Future Work below).

IV. Empirical Evaluation

The utility of **petal** is demonstrated with an application of an Illumina RNA-seq whole-genome sequencing experiment of the mountain pine beetle (*Dendroctonus ponderosae*). Mountain pine beetles are obligate parasites of pine trees. They have destroyed a wide area of forest land and are a serious threat to conifer forests in the western North America. They rely on aggregation pheromones to coordinate the "mass attacks" necessary to overwhelm a host tree's defenses and thus successfully colonize a tree. A molecular level understanding of this process may provide new methods to manage these devastating pests. Although pheromone biosynthetic pathways have been previously studied, the enzymes involved have not yet been completely identified, characterized, and understood [36, 37, 38]. Aw *et al.* presented the first genomic analysis of the mountain pine beetle and identified candidate genes encoding enzymes involved in pheromone-biosynthesis by studying their gene expression patterns [36], which yielded two confirmed pheromone-biosynthesizing enzymes [39]. The hypothesis is that genes encoding these enzymes are coordinately regulated. Of particular interest is a group of 28 genes previously implicated in pheromone biosynthetic pathways.

A. Data

In this experiment, the Illumina NextSeq 500 platform was used to generate RNA-seq measures of gene expression measures of more than 13,000 genes of the mountain pine beetle. Four biological replicates were collected for each of the four specimen types: fed/unfed male/female. Sequences were trimmed and filtered for nucleotide-base quality and 19-35 million sequences were aligned to the *Dendroctus ponderosae* reference genome. Unambiguously aligned sequences were counted for all annotated mountain pine beetle genes. Count data underwent standard protocols for low-count filtering, upper quartile normalization and transformation into counts per million following the DESeq2 processing pipeline [40]. Experimental findings relevant to beetle biology and biochemistry will be described in a forthcoming manuscript, in which the data will be made publicly available.

B. Application of Tool

After data quality control, the dataset contained 11,342 gene identifiers across 16 measures, which was uploaded into **petal**. Upon upload, the **petal** histogram clearly showed that the data were distributed non-normally, thus the similarity metric was left as the non-parametric default (SP). The goal is to generate a scale-free small-world network model on the expression data as quickly as possible; **petal** is left to select the correlation metric threshold automatically. The list of 28 gene identifiers of interest were uploaded and selected to be analyzed together as they have been hypothesized to play a joint role in the in pheromone biosynthetic pathways. The **petal** run was performed on a server with two 2.5GHz processors and 256GB RAM, and took 4.33 hours utilizing at most 7.5GB RAM at any time.

C. Results

A series of seven thresholds ranging between 0.956 and 0.734 was determined based on SP measures of all pair-wise comparisons to find a scale-free and small-world network. For all seven thresholds the adjacency matrices were generated and their topological properties calculated and presented in the NetworkStats.txt file (Fig. 3). Properties in Figure 3 are used by **petal** to identify the 'best' threshold as explained in III.A.2. The first column is the list of considered thresholds. The second and third columns represent the values obtained from the linear regression on the log-transformed degree distribution, meanCC is the mean cluster coefficient, meanPath is the average path length between vertex pairs, %used indicates the percentage of genes used from the original dataset signifying how many genes have connections within the specific network model, and %bigComp describes how many of the network's vertices are within the biggest component. **petal** identified a SP threshold of 0.808 to produce the 'best' scale-free, small-world network model, that is used for downstream analysis as described in *III.A.4*. Inspecting Figure 3, we see that thresholds above 0.845 are excluded from the decision process for 'best' threshold as the biggest component of those networks include less than 95% of the network's vertices. Also thresholds 0.771 and 0.734 are excluded due to their low coefficient of determination (R^2). Consequently, only 0.808 and 0.845 remain, the network based on 0.808 contains about 700 more genes than the model based on 0.845, providing a better whole-systems approach. As a result 0.808 is set to the 'best' network model and 0.845 is an alternative model.

threshold	R^2	slope/power	meanCC	meanPath	%used	%bigComp
0.956	0.8418	-1.7148	0.4408	6.8906	20.9751	22.0681
0.919	0.8974	-1.6175	0.3703	11.1274	49.5415	84.9617
0.882	0.8899	-1.4481	0.3787	7.1925	72.2448	93.5074
0.845	0.8623	-1.2386	0.3953	5.6573	85.6727	96.573
0.808	0.8191	-1.0467	0.4183	4.7138	93.8106	98.656
0.771	0.7666	-0.9265	0.4407	4.0383	98.1132	99.5237
0.734	0.7103	-0.8506	0.4654	3.5486	99.4886	99.9025

Fig. 3: Screen shot of output file – `NetworkStats.txt`.

There are 13 vicinity networks (VNs) obtained from the list of 28 genes of interest. Two of them are of special interest as they contain 5 and 6 of the 28 genes, with a density of 88% and 89%, respectively. These two VNs overlap in 4 out of the 28 genes of interest; overall these two VNs have a total intersection of 28 genes. The subnetwork of the 24 neighbor genes and the 7 genes of interest resulted in a subnetwork with a density of 98.71%. Within this 31 gene subnetwork 6 edges are missing to be a clique, resulting in a fuzzy clique. The profiles of these 31 genes indicate higher expression in male than in female mountain pine beetles as seen in Figure 4. The expression difference is much more dramatic in the males which have not yet infested a tree and therefore have not eaten. This fuzzy clique is scientifically notable because some of the members encode enzymes with activities that are predicted to catalyze uncharacterized steps of synthesis in the pheromone component. A closer evaluation of the genes' functions within the 31-node fuzzy clique identifies genes that encode enzymes

already confirmed as pheromone biosynthesizers. In addition, this fuzzy clique includes genes which previously have been predicted to catalyze known steps in the pheromone biosynthetic pathway. Within this identified grouping, the scientist is now able to narrow down targets for further wet-lab examinations.

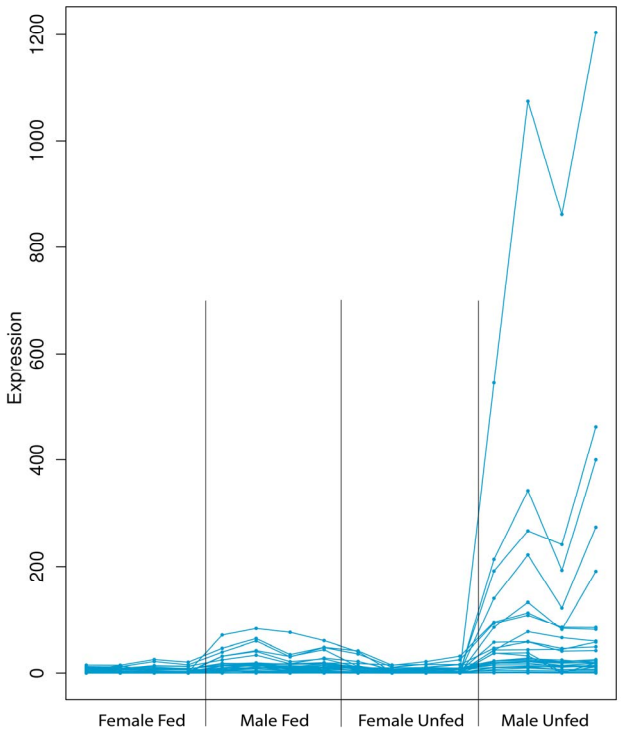

Fig. 4: 31 gene expression profiles belonging to a subnetwork with density of 98.71%.

Another interesting subnetwork is a vicinity network obtained from four other genes of interest. All 85 common neighbors are tightly interconnected: the 89 gene subnetwork has a density of 96%. Biologically, this subnetwork presents a group of 89 very similarly expressed genes, including various cytochromes P450; this grouping agrees with a hypothesized link between tree resin detoxification and pheromone production [36, 39].

Overall, this approach enables the researcher to quickly view genes with highly similar expression patterns. With current annotation of the genes at hand, simple observations of the similarity or differences of functions of similarly-behaving genes can be made (Fig. 4).

V. Conclusions And Future Work

Here, we have presented a novel tool to construct co-expression networks, implemented in **R**. Its innovation lies in 1) a choice of metric statistically appropriate for the data via visual inspection of data distribution, 2) generating only biologically meaningful network models by automatically enforcing the properties of scale-free and small-world, and 3) the ability to analyze whole-system data, rather than small, pre-selected subsets (e.g., DE genes). **petal** takes advantages of graph-theoretical approaches, appropriate statistics, and mathematical precision to produce unbiased and biologically meaningful results. The computation of a biologically meaningful metric cutoff threshold is one of the innovative approaches of **petal**: this fundamental component of the network model is computed objectively to result in biological meaning, rather than by subjective user selection.

petal produces network models that present associations among genes of a studied system based on experimental data. These models provide a comprehensive view of the entire system which comes at a cost of longer computational runtime compared to most other current tools (e.g., WGCNA). On the other hand, user time is drastically reduced due to restricting user-intervention, decreasing the manual execution of computational steps. WGCNA, although very low in computational costs, does not purposefully generate small-world networks, and ensures scale-free networks only with user intervention. Cytoscape [30, 35] is a very popular tool that is used to view networks; the construction of co-expression networks is unique to one plugin only allowing the PCC metric. **petal** specifically generates biologically meaningful co-expression networks based on metrics appropriate for the data, and allows the user to examine densely connected subnetworks of genes of interest, both mathematically, and via an additional viewer such as Cytoscape or Pajek. The tool is very user-friendly by requiring little prior knowledge of network science without sacrificing the quality output that comes from complex well graph-theoretically defined networks. **petal**'s adaptability allows for the analysis of experimental expression data of most sizes.

The next steps in **petal**'s development are to make the **petal** GUI publicly available and formally construct an **R**-library for submission to **R** or Bioconductor. Additionally, we plan to parallelize **petal** to decrease computing time. Current empirical tests of ice plant gene expression datasets of 5000 genes across seven conditions yielded networks for five thresholds in 1.35 hours (PCC metric) and 2.07 hours (SP) on a server with two 2.5GHz processors and 256GB RAM, using a maximum of 1GB RAM at any given time. Parallelization will allow the inclusion of additional metrics to define association between vertices, making our approach attractive to any researchers who are interested in constructing large-scale, small-world, scale-free networks. More specific to the life scientists, to strengthen the downstream analysis of **petal** we are developing the integration of a user-provided annotation file for easy identification of over-represented gene groups within VNs.

The source code can be found at:
https://github.com/julipetal/petalNet.

Acknowledgment

This work is based upon work supported by the Department of Energy (DOE), Office of Science, Genomic Science Program under Award Number DE-SC0008834. This work was also made possible by a grant from the National Institute of General Medical Sciences (P20GM103440) from the National Institutes of Health through its support of the Nevada Center for Bioinformatics. The contents of this manuscript are solely the responsibility of the authors and do not necessarily

represent the official views of the DOE or the NIH. We thank Tyler C. Sorey for the initial setup of the tool's graphical user interface and Sebastian Smith for his continuing support of **petal**'s development. We also express our gratitude to Claus Tittiger and Jeff Nadeau for allowing us to apply **petal** to their current mountain pine beetle experiment. Lastly, we appreciate the constructive criticism provided by all reviewers.

REFERENCES

[1] B. Schwikowski, P. Uetz, and S. Fields, "A network of protein-protein interactions in yeast", *Nature Biotechnology*, 18(12):1257-1261, 2000.

[2] M.J. Jeger, M. Pautasso, O. Holdenrieder, and M.W. Shaw, "Modelling disease spread and control in networks: implications for plant sciences", *New Pytologist*, 174(2):279-297, 2007.

[3] Y. Chen, J. Zhu, P.Y. Lum, X. Yang , S. Pinto, D.J. MacNeil, *et al*., "Variations in DNA elucidate molecular networks that cause disease", *Nature* 452: 429–435, 2008.

[4] R. De Smet and K. Marchal, "Advantages and limitations of current network inference methods", *Nature Reviews. Microbiology*, 8(10), 717-729, 2010.

[5] M. Maschietto, A.C. Tahira, R. Puga, L. Lima, D. Mariani, S. Paulsen Bda, *et al*., "Co-expression network of neural-differentiation genes shows specific pattern in schizophrenia", *BMC Medical Genomics*, 16(8:23), 2015.

[6] S. Horvath and J. Dong, "Geometric interpretation of gene coexpression network analysis", *PLoS Computational Biology*, 4(8), e1000117, 2008.

[7] X. Xiao, A. Moreno-Moral, M. Rotival, L. Bottolo, and E. Petretto, "Multi-tissue analysis of co-expression networks by higher-order generalized singular value decomposition identifies functionally coherent transcriptional modules", *PLoS Genetics*, 10(1):e1004006, 2014.

[8] S. de Jong, M.P. Boks, T.F. Fuller, E. Strengman, E. Janson, C.G. de Kovel, *et al.*, "A gene co-expression network in whole blood of schizophrenia patients is independent of antipsychotic-use and enriched for brain-expressed genes", *PLoS One*, 7(6):e39498, 2012.

[9] G. Bidkhori, Z.Narimani, S. Hosseini Ashtiani, A. Moeini, A. Nowzari-Dalin, and A. Masoudi-Nejad, "Reconstruction of an integrated genome-scale co-expression network reveals key modules involved in lung adenocarcinoma", *PloS ONE*, 8(7), e67552, 2013.

[10] G. Broderick, J. Fuite, A. Kreitz, S.D. Vernon, N. Klimas, and M.A. Fletcher, "A formal analysis of cytokine networks in chronic fatigue syndrome", *Brain, Behavior, and Immunity*, 24(7), 1209–1217, 2010.

[11] B. Munneke, K.A. Schlauch, K.L. Simonsen, W.D. Beavis, and R.W. Doerge, "Adding confidence to gene expression clustering", *Genetics*, 170(4), 2003–2011, 2005.

[12] L. Song, P. Langfelder, and S. Horvath, "Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics*, 13:328, 2012.

[13] J.C. Cushman, R.T. Tillett, J.A. Wood, J.M. Branco, and K.A. Schlauch, "Large-scale mRNA expression profiling in the common ice plant, Mesembryanthemum crystallinum, performing C3 photosynthesis and Crassulacean acid metabolism (CAM)", *Journal of Experimental Botany*, 59(7), 1875–1894, 2008.

[14] M. Dehmer, F. Emmert-Streib, A. Graber, and A. Salvador, "Applied statistics for applied biology, methods in systems biology":, Wiley-Blackwell, Germany, 2011.

[15] B. Zhang and S. Horvath, "A general framework for weighted gene co-expression network analysis", *Statistical Applications in Genetics and Molecular Biology*, 4, Article17, 2005.

[16] M. E. J. Newman, "Networks: An Introduction", New York, NY: Oxford University Press, 2012.

[17] A.L. Barabási and R. Albert, "Emergence of scaling in random networks", *Science*, 286(5439), 509-512, 1999.

[18] D.J. Watts and S.H. Strogatz, "Collective dynamics of 'small-world' networks", *Nature*, 393(6684), 440-442, 1998.

[19] V. van Noort, B. Snel, and M.A. Huynen, "The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model", *EMBO Reports*, 5(3), 280-284, 2004.

[20] J. Ruan, A.K. Dean, and W. Zhang, "A general co-expression network-based approach to gene expression analysis: comparison and applications", *BMC Systems Biology*, 4:8, 2010.

[21] Xulvi-Brunet R, Li H, "Co-expression networks: graph properties and topological comparisons. *Bioinformatics*, 26(2), 205-214, 2010.

[22] A.L. Barabási and Z.N. Oltvai, "Network biology: understanding the cell's functional organization", *Nature Reviews, Genetics*, 5(2), 101-113, 2004

[23] A.L. Barabási, "Scale-free networks: a decade and beyond", *Science*, 325(5939), 412-413, 2009.

[24] D.J. Watts, "Six degrees: the science of a connected age", WW Norton & Company, Inc. New York, 2004.

[25] J. Ruan and W. Zhang, "Identification and evaluation of weak community structures in networks", Proceedings of the Twenty-First National Conference on Artificial Intelligence. pp. 470–475, 2006.

[26] P. Langfelder and S. Horvath, "WGCNA: an R package for weighted correlation network analysis", *BMC Bioinformatics*, 9:559, 2008.

[27] C. Soneson and M. Delorenzi, "A comparison of methods for differential expression analysis of RNA-seq data", *BMC Bioinformatics*, 9, 14:91, 2013.

[28] V. Srinivasasainagendra, G.P. Page, T. Mehta, I. Coulibaly, and A.E. Loraine, "CressExpress: a tool for large-scale mining of expression data from Arabidopsis", *Plant Physiology*, 147(3):1004-1016, 2008.

[29] Y. Ogata, H. Suzuki, N. Sakurai, and D. Shibata, "CoP: a database for characterizing co-expressed gene modules with biological information in plants", *Bioinformatics (Oxford, England)*, 26(9), 1267–1268, 2010.

[30] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, *et al*., "Cytoscape: a software environment for integrated models of biomolecular interaction networks", *Genome Research*, 13(11), 2003.

[31] V. Batagelj, Pajek, http://mrvar.fdv.uni-lj.si/pajek/ (accessed July 2015)

[32] A. Ben-Dor, R. Shamir, and Z. Yakhini, "Clustering gene expression patterns", *Journal of Computational Biology*, 6(3/4), 281–297, 1999.

[33] R. Ihaka and R. Gentleman, **R** (programming language) http://cran.r-project.org/ (accessed July 2015)

[34] C. Gabor, igraph **R**-library, http://igraph.org/redirect.html (accessed July 2015)

[35] R. Saito, M.E. Smoot, K. Ono, J. Ruscheinski, P.L. Wang, S. Lotia, *et al.,* "A travel guide to Cytoscape plugins", *Nature Methods*, 9(11), 1069-1076, 2012.

[36] T. Aw, K. Schlauch, C.I. Keeling, S. Young, J.C. Bearfield, G.J. Blomquist, and C. Tittiger, "Functional genomics of mountain pine beetle (*Dendroctonus ponderosae*) midguts and fat bodies", *BMC Genomics*, 11(1):215, 2010.

[37] C.I. Keeling, M.M. Yuen, N.Y. Liao, T.R. Docking, S.K. Chan, G.A. Taylor, *et al.*, "Draft genome of the mountain pine beetle, *Dendroctonus ponderosae Hopkins*, a major forest pest", *Genome Bioliology*, 14(3):R27, 2013.

[38] C.I. Keeling, H. Henderson, M. Li, M. Yuen, E.L. Clark, J.D. Fraser, *et al.,* "Transcriptome and full-length cDNA resources for the mountain pine beetle, *Dendroctonus ponderosae* Hopkins, a major insect pest of pine forests", *Insect Biochemistry and Molecular Biology*, 42(8), 525-536, 2012.

[39] M. Song, P. Delaplain, T.T. Nguyen, X. Liu, L. Wickenberg, C. Jeffrey, *et al*., "exo-Brevicomin biosynthetic pathway enzymes from the Mountain Pine Beetle, *Dendroctonus ponderosae*", *Insect Biochemistry and Molecular Biology*, 53, 73-80, 2014.

[40] M.I. Love, Huber W, and Anders S., "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2", Genome Biology, 15(12):550, 2014.