(Differential) Co-Expression Analysis of Gene Expression: A Survey of Best Practices

Hussain Ahmed Chowdhury, Dhruba Kumar Bhattacharyya, and Jugal Kumar Kalita

Abstract—Analysis of gene expression data is widely used in transcriptomic studies to understand functions of molecules inside a cell and interactions among molecules. Differential co-expression analysis studies diseases and phenotypic variations by finding modules of genes whose co-expression patterns vary across conditions. We review the best practices in gene expression data analysis in terms of analysis of (differential) co-expression, co-expression network, differential networking, and differential connectivity considering both microarray and RNA-seq data along with comparisons. We highlight hurdles in RNA-seq data analysis using methods developed for microarrays. We include discussion of necessary tools for gene expression analysis throughout the paper. In addition, we shed light on scRNA-seq data analysis by including preprocessing and scRNA-seq in co-expression analysis along with useful tools specific to scRNA-seq. To get insights, biological interpretation and functional profiling is included. Finally, we provide guidelines for the analyst, along with research issues and challenges which should be addressed.

Index Terms—Co-expression analysis, co-expression network, differential co-expression analysis, disease gene prediction, gene regulatory network, differential networking, differential connectivity, RNA-seq, microarray, scRNA-seq, gene expression

1 Introduction

A trascriptomic study attempts to understand the different functions of molecules inside a cell and interactions among them by analyzing gene expressions. Analysis of gene expression data is an active research area. Gene expression measures the amount of mRNA produced by a gene at a particular time. Two widely used technologies to generate gene expression data are microarrays and RNA-seq (bulk RNA-seq). These technologies facilitate in-depth transcriptomic studies, and can estimate gene expression of millions of cells in a single experiment. Generally, gene expression reflects the average expression values of expressions across cells. The average is a gross statistics and may not be a good representation of the distribution of gene expression levels. Latest advances in RNA-seq has led to the development of scRNA-seq (single cell RNA-seq), which can address the above issue. The main difference between scRNA-seq from RNA-seq is that the sequencing library contains a single cell. Like microarrays and RNA-seq, scRNA-seq data analysis requires preprocessing for effective downstream analysis. Some widely used preprocessing steps include removal of genes with low count, imputation, batch effect removal, normalization, and transformation.

There are many data analysis strategies available to understand the function of a gene in a systematic way. Gene Co-expression Network analysis is commonly used to predict gene function at a genome level. It is also a way to discover functions of unknown genes and their associations with diseases. Co-expression networks for microarray and RNA-seq data have many commonalities and differences. Co-expression analysis of scRNA-seq data can help solve biological problems both at the cell as well as at the gene levels.

Co-expression analysis finds genes that have the tendency to exhibit similar expressions in a group of samples, and its types include module finding, hub gene finding, guilt by association analysis, regulatory network construction integrative analysis and Preservation and Disruption Analysis. Differential co-expression analysis finds modules of genes whose co-expressions significantly vary across conditions. This analysis can uncover biologically interesting results, which cannot be found by differential expression and co-expression analysis alone. A pair of genes may show shift-and/or-scale, cross/inverse and rewiring expression patterns across different sample groups. In case of a group of genes, co-expression structures across sample groups may be different. The main aim of differential network and differential connectivity analysis is to find a sub-network or a set of connected genes that operates abnormally and affects a significant number of neighbors across sample groups. For biomarker identification, the assumption is that the genes which change behavior with respect to a significant number of neighbors across conditions are more likely to be interesting [1]. Hybrid analysis combines several analysis approaches to better understand complex disease mechanisms. Functional profiling is also an important step in transcriptomic studies to characterize molecular functions or pathways in which interesting genes from different analysis are involved.

- *H.A. Chowdhury and D.K. Bhattacharyya are with the Department of Computer Science and Engineering, Tezpur University, Sonitpur, Assam 784028, India. E-mail: hachowdhury09@yahoo.in, dkb@tezu.ernet.in.*
- *J.K. Kalita is with the University of Colorado, Colorado Springs, CO 80933-7150. E-mail: jkalita@uccs.edu.*

Manuscript received 20 June 2018; revised 4 Dec. 2018; accepted 8 Jan. 2019. Date of publication 16 Jan. 2019; date of current version 6 Aug. 2020. (Corresponding author: Dhruba Kumar Bhattacharyya.) Digital Object Identifier no. 10.1109/TCBB.2019.2893170

1545-5963 © 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Universidad Nacional Autonoma De Mexico (UNAM). Downloaded on September 18,2024 at 20:53:55 UTC from IEEE Xplore. Restrictions apply.

1.1 Prior Surveys and Motivations

There are a few published surveys such as Van Dan et al. [2] and Aoki et al.[3] on (differential) co-expression analysis of genes. Van Dam et al. provide an overview of co-expression analysis and present tools and methods to create co-expression networks from genes expression and analyze such networks. This recent review also explains how these methods and tools are useful in identifying regulatory genes for a disease. Integration of omic data with co-expression networks is also discussed along with the future prospect of co-expression analysis. Aoki et al. discuss all aspects of co-expression analysis such as network architecture, protocol, module detection and finding of new genes, but their discussion is limited to microarray data only [3]. Most of the content of [3] is covered by [2], but, in [2], discussions about differential connectivity and differential networking are not included, and no comparison of any kind is given. Van Dam et al. [2] discuss differential co-expression analysis in a very limited way, which is not sufficient. In addition, neither survey includes preprocessing, hybrid analysis and scRNA-seq data analysis. Any recommendation for practitioners is also missing in both reviews. Though Van Dam et al. include a future perspective on co-expression analysis, it would be fitting to include research issues and challenges as well. The following points make our survey differ from previous surveys.

a) Like [2] and [3], we discuss co-expression network analysis along with useful tools. In addition, we discuss differential co-expression analysis more extensively by including differential co-expression patterns and detection strategies.
b) Unlike [2] and [3], our survey is not restricted to discussing only preprocessing, and co-expression analysis of scRNA-seq data. Moreover, we also discuss differential co-expression analysis tools which can be adapted for scRNA-seq data.
c) Unlike [2] and [3], we discuss preservation and disruption analysis.
d) Like [2], we do not restrict ourselves to a discussion of integrative analysis of co-expression networks. We include hybrid analysis along with tools useful for hybrid analysis.
e) Unlike other surveys, we include a discussion on network inference to reconstruct transcriptional networks.
f) Unlike other surveys, we discuss differential networking along with comparison of differential expression, differential networking, differential co-expression and differential connectivity.
g) Unlike [3], we present many necessary and useful tools throughout the paper.
h) Unlike the above mentioned surveys, we include biological interpretation and functional analysis.

1.2 Our Contributions

The major contributions of this survey are stated below.

- A detailed discussion of co-expression network and co-expression analysis and their types.
- A discussion of co-expression networks in RNA-seq versus microarrays and roles of RNA-seq in co-expression analysis.
- A discussion on co-expression analysis of scRNA-seq data.
- A organized discussion of differential co-expression analysis along with different patterns and detection strategies and comparison of its variants. Hybrid analysis is also included.
- A large number useful tools for gene expressions analysis are presented throughout the paper.
- A discussion of biological interpretation and functional analysis.
- A set of useful guidelines for the data analyst is provided.
- Important research issues and challenges from both theoretical and practical viewpoints are included.

2 Preprocessing of Gene Expression Data

RNA-seq and Microarray technologies obtain average gene expression values of cells, whereas, in scRNA-seq, gene expressions can be quantified at a cell level in an experiment. These three technologies require different types of preprocessing before performing downstream analysis such as differential expression analysis, co-expression analysis or differential co-expression analysis. Some widely used preprocessing techniques for both RNA-seq and scRNA-seq are discussed in the following sections.

2.1 Elimination of Low Read Counts

The presence of a large abundance of zeroes in the count data matrix creates problems during any kind of analysis. Even a perfect correlation between a pair of genes may not be detectable in the presence of a large number of zeroes. In addition, some genes contain very low numbers of mapped reads, which may be mapped randomly during alignment without providing much information for downstream analysis and biasing the results. So, discarding entries with low read counts is recommended [4].

2.2 Imputation

RNA-seq and scRNA-seq datasets generally contain large numbers of zeroes, i.e., zero inflation. So, one needs to have a strategy for treating it. scRNA-seq data also suffer from dropout events, which may mislead downstream analysis. To address such problems, imputation is required. Several imputation techniques have been proposed for microarray data. These techniques cannot be directly applied on RNA-seq and scRNA-seq data because of the discrete distribution of count data. In addition, scRNA-seq quantifies gene expression profile of a single cell, and does not average like other technologies working with gene expression data [5]. So, special care is needed to address this issue. Some well-known imputation methods for scRNA-seq are scUnif [6], MAGIC [7], scImpute, DrImpute [8], SAVER [9], and BISCUIT [10]. scUnif and MAGIC can be used for RNA-seq data imputation as well. A detailed discussion of some of these tools and their comparisons are reported in [11]

2.3 Batch Effect Removal

Batch effects in RNA-seq are mostly due to the measurements, which are influenced by laboratory conditions. These common and powerful sources of variation are eliminated using appropriate preprocessing by considering global properties of data. COMBAT [12] and ARSyN [13] are two

Authorized licensed use limited to: Universidad Nacional Autonoma De Mexico (UNAM). Downloaded on September 18,2024 at 20:53:55 UTC from IEEE Xplore. Restrictions apply.

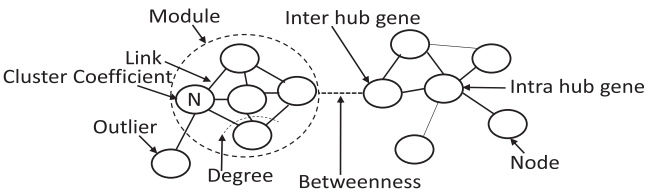

Fig. 1. Co-expression network architecture. Degree: Total number of links incident on a node. Clustering coefficient of node N: Ratio between the number of links present in N's neighbors and all possible links in N's neighbor. Betweenness: The total number of possible paths through a node/link between pairs of nodes.

methods used in removing batch effects from RNA-seq as well as microarray data [14].

2.4 Normalization

Normalization makes gene expression levels comparable across different samples and genes to make valid inferences. The aim of normalization is to eliminate biases that arise due to the biological differences of interest. For effective normalization, one must consider factors such as GC-content, transcript size, sequencing depth, and insert size [15]. The NVT package can identify the best normalization approach for a RNA-seq dataset by analyzing from 11 existing normalization methods [16]. Zyprych-Walczak et al. provide a procedure to find a suitable normalization method for a specific dataset [17]. Some popular methods for RNA-seq normalization include counts per million (CPM), RPKM [18], TPM [19], UQ [20], TMM [21] and DESeq [22]. Vallejos et al. performed a study on the suitability of RNA-seq based methods for scRNA-seq data and found that they can mislead downstream analysis. They also found that RNA-seq normalization methods may have adverse consequences in subsequent analysis of scRNA-seq, e.g., in the detection of highly variable genes.[23]. So, scRNA-seq specific normalization methods are required. In scRNA-seq data normalization, global scaling, which removes cell-specific systematic biases is the most widely used strategy. This is done by scaling gene expressions by a constant factor within each cell [23]. Some other methods specific to scRNA-seq data are GRM [24], SAMstrt [25] and SCnorm [26].

2.5 Transformation

RNA-seq data are discrete measurements of gene expression. Therefore, transformation is required to use statistical methods developed for data that follow normal distribution. Transformation techniques include log transformation, variance-stabilizing transformation (VST), rank-based transformation, and Box-Cox transformation [27]. These transformations may be used to transform discrete RNA-seq data to normal distribution as in the case of microarrays [28]. Zwiener et al. performed a simulation study of transformation techniques and found that transformations potentially improve the detection of important genes [27]. Log transformation and VST are widely used in RNA-seq and scRNA-seq data [29], [30]. Log transformation cannot transform data to exact normal-like distribution, but, this method is still being used [28]. VST [22], implemented in the DESeq tool [4], can be used to obtain covariates with variances. This transformation has fewer skewed values, but still can include large values that are very different from the mean value.

3 Co-Expression Network

A co-expression network enables researchers to identify groups of tightly correlated genes associated with biological processes or pathways, and it illustrates the correlation between genes in a graphical form. A gene co-expression network generally has a scale-free topology [31]. Network analysis extracts complex relationships between the genotype and phenotype, and reveals the biological mechanisms of complex diseases [32], [33]. The architecture of a co-expression network is given in Fig. 1. Steps for co-expression network construction and analysis are as follows and a graphical representation is presented in Fig. 2. After preprocessing of gene expression data, pairwise correlation computation is performed between each possible pair of genes. Then, a network is constructed where nodes represent genes and edges represent computed pairwise correlation between pairs of co-expressed genes. Next, clustering is performed to find co-expressed genes groups, i.e., modules. The found modules can undergo various types co-expression analysis discussed in Section 4. Finally, functional profiling, comprising pathway and enrichment analysis, is performed to obtain biological insights present in the list of genes.

Some widely used correlation measures to compute pairwise correlation are Pearson's correlation, Spearman's correlation, and mutual information. Kumari et al. compared eight gene association methods and their effects on co-expression network construction and biological knowledge discovery [34]. They reported that effectiveness of these methods depends on the data properties and the biological processes. The main challenge in co-expression analysis is to choose the appropriate algorithm because different clustering algorithms find different groups of genes as co-expressed and influence further analysis. In addition, clustering results also influence by proximity measures such as statistical or semantic similarity or combination of both.

Our experimental study with a large number of algorithms reveals that with the variations in proximity measures, the outcomes of supervised or unsupervised learning algorithm varies significantly. Our observation is that it is mostly due to the use of statistical proximity measures for expression data such as euclidean distance or Pearsons correlation or their variants. But, the use of semantic measures (for categorical data) such as [35]: Resink, Lin, and Wang can help minimize the inconsistency to a great extent. There are good examples of using weighted combination of both these types of measures also [36], [37], [38]. However, we also feel that it is a challenging task to achieve a single biological measure that can capture both numeric expression proximity as well as categorical proximity (GO based) meaningfully to ensure a consistent performance across all gene expression data. Systematic noise may hinder in finding the true co-expressions of genes. RUVcorr is package to deal with this a problem by removing noise from the data more accurately [39].

3.1 Types of Co-Expression Networks

A co-expression network can be constructed in different ways based on the way correlation is recorded between pairs of genes. Co-expression networks can be broadly divided into three types as follows.

Unsigned Co-Expression Networks. Many correlation measures return correlation values between a pair of genes in the range form -1 to $+1$. In an unsigned co-expression

Authorized licensed use limited to: Universidad Nacional Autonoma De Mexico (UNAM). Downloaded on September 18,2024 at 20:53:55 UTC from IEEE Xplore. Restrictions apply.

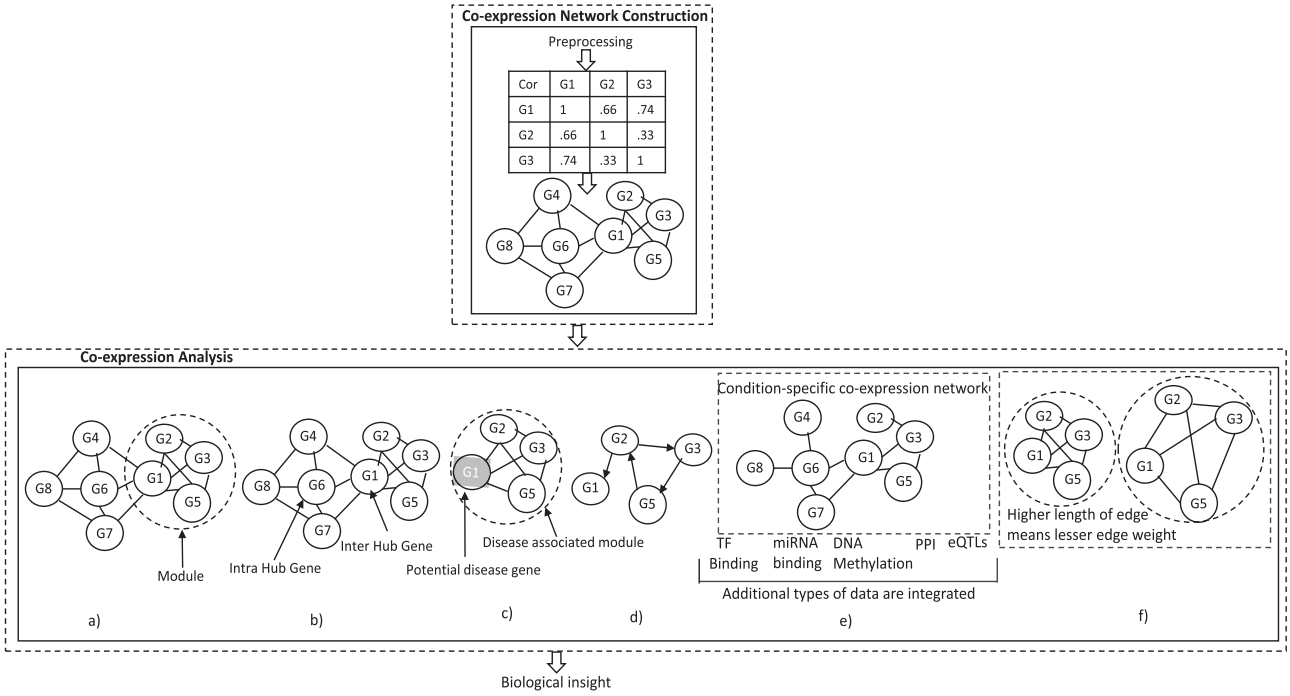

Fig. 2. Construction of co-expression network and co-expression analysis. In co-expression network construction, after preprocessing, pairwise correlations are calculated and then network construction is performed. In co-expression analyses, different analysis such as a) clustering to find modules, b) hub genes finding, c) guilt by association, d) regulatory network construction e) integrative analysis, and f) preservation and disruption analysis, are performed. Functional profiling is required to obtain biological insights.

network, absolute values of correlation measures are used, i.e., a negatively correlated pair of genes is also considered co-expressed because these genes may be positively co-regulate other sets of genes. However, such pairs may disrupt the structure of the co-expression networks. This problem is handled in the following types of networks.

Signed Co-Expression Networks. Here, the correlation values between a pair of genes are transformed to range from 0 (perfectly negative correlation) to 1 (perfect positive correlation). A correlation value <0.5 is considered negatively correlated and a correlation value >0.5 is considered positively correlated.

Weighted and Un-Weighted Co-Expression Networks. In weighted co-expression networks, all nodes in the network are connected to each other like complete networks, and continuous weight ranges from 0 to 1 are assigned to each edge based on the tendency of a pair of genes to be co-expressed. Unweighted co-expression networks are also similar to weighted co-expression networks. The only difference is that the weights assigned to the edges are either 1 or 0, obtained by transforming (also by comparing with user threshold) the edge weights. Here, 1 and 0 are used to represent whether a pair of genes is connected or disconnected.

Generally, microRNAs and transcription factors express their functions through regulation of gene expression. So, to explore their functions in gene expression data, signed co-expression network is more useful. To find modules responsible for phenotype variations or biological processes, a signed co-expression network can also be used because in such networks, modularity is high, allowing separation of biologically meaningful modules. Weighted co-expression networks are more robust in producing biologically meaningful results. These networks are widely used in module centric analysis such as hub gene finding, guilt by association analysis and tissue-specific analysis. These may be used to predict new biomarker(s) and for disease sub-type identification.

3.2 Co-Expression Network in RNA-Seq versus Microarray

To analyze traditional microarray data, there are many similarity measures. However, which of among them are suitable as similarity measures for RNA-seq data (without transforming) is yet to be established [40]. Jin et al. performed a comparative study on breast cancer gene expression data from microarrays and RNA-Seq, and found a decent correlation among gene expressions generated from the two platforms [41]. Iancu et al. performed a comparative study of co-expression networks from both RNA-seq and microarray data using Pearson correlation coefficient [42]. They saw a large amount of overlap between the networks constructed from these two types of data. They also found high connectivity, centrality and heterogeneity in RNA-seq networks. Similarly, Giorgi et al. observed a decent similarity between the networks constructed from both RNA-seq and microarray data [43]. However, RNA-seq data show higher sensibility to VST. Sekhon et al. performed a comparative study of RNA-seq- and microarray-based gene expression data for transcriptome analysis [44]. They found that tissue specific gene expression can be captured by RNA-seq very well. But, they also found similar types of co-expression networks for both technologies. Ballouz et al. found that microarray and RNA-seq derived co-expression networks are very similar in terms of functional connectivity (the amounts of functional involvement of a pair of genes) [36] and are very dissimilar in terms of overall network topology [45].

3.3 RNA-Seq in Co-Expression Analysis

One major advantage of RNA-seq is that it can measure almost 70,000 non-coding RNAs, which is impossible with

Authorized licensed use limited to: Universidad Nacional Autonoma De Mexico (UNAM). Downloaded on September 18,2024 at 20:53:55 UTC from IEEE Xplore. Restrictions apply.

microarrays. As a result, RNA-seq plays a vital role in finding disease mechanisms [2]. During analysis, consideration of non-coding RNAs gives a better understanding of regulatory mechanisms and biological processes. Co-expression analysis of RNA-seq can be used to focus on non-coding RNAs and splice variants [46], [47].

Co-expression analysis on a splice variant is challenging when many splice variants are associated with same exon because we cannot predict which splice variant is expressed [47], [48], [49]. This leads to biases in results. In RNA-seq, it is possible to perform expression quantification at the exon level, gene level and transcript level [50]. Exon level quantification can be used to construct co-splicing networks [51], whereas, in gene co-expression networks, the co-expression signal suffers from biases because expression of different transcripts originating from the same gene are generally aggregated [51]. Co-splicing networks do not suffer from this issue. In RNA-seq co-expression analysis, two genes are said to be co-expressed if their distinct splice variants show coordinated expression. This approach has the ability to identify novel and unique functional modules, which are not found in traditional co-expression analysis [52].

To perform gene level co-expression analysis, overlapping isoforms are merged, and then constructing a co-expression network is the most common and effective way. The drawback is that there is a chance of losing information about transcripts coming from the same gene. A transcript or exon level co-expression network can be constructed, but the size of the network will be huge, leading to the need for high computational resources, making it a challenge. One way, this challenge may be handled is by performing distributed co-expression analysis i.e., using high performance computing facilities in co-expression analysis. For example, the WGCNA package has the ability to use multiple cores during co-expression analysis.

Determining the optimal sample size and read depth required for co-expression analysis of RNA-seq is challenging. Ballouz et al. [45] suggested that one consider at least 20 samples in co-expression analysis of RNA-seq. More samples generally provide more functional connectivity in the co-expression network. So, it is essential to have a number of samples for co-expression analysis. In the same way, read depth is another factor which can influence co-expression analysis. Higher read depth is always recommended because higher read depth increases the accuracy of expression measurement. More than 10 million reads per sample are ideal for co-expression analysis [45]. Co-expression analysis also depends on the quality of the annotated genome file. For better quality genome data, higher read depth is preferable. Ballouz et al. use 3 different pipelines to generate the RNA-seq count data and perform co-expression analysis [45]. They found that co-expression measures have impact on network topology, but not on functional connectivity. Petal is a tool for co-expression network modeling for whole genome experiments and provides meaningful results for further biological interpretation [53]. It does not assume a data distribution, making it useful for RNA-seq data analysis as well. Large-Scale Transcriptome Analysis Pipeline (LSTrAP) preprocesses RNA-seq data and produces co-expression networks, along with comparative genomic data to enable effective downstream analysis [54]. The output of LSTrAP is compatible with Cytoscape [55], thus enabling the mining of networks for novel biological information.

3.4 scRNA-Seq in Co-Expression Analysis and Its Application

There are numerous applications of scRNA-seq because it enables researchers to study gene expression at a new level which was not possible using bulk RNA-seq. Some co-expression based applications [56] are cellular state and cell type identification, where clustering of cells is performed using count data to find hidden tissue heterogeneity [57]; co-regulated network and gene regulatory network (GRN) identification; and characterization of diversity in transcription between individual cells, where samples in the count data are replaced with cells before analysis and then, like RNA-seq, we can identify co-regulated genes.

In scRNA-seq, gene-level analysis includes co-expression, differential expression and gene variability analysis. In principle, samples in the count data are replaced with cells before co-expression analysis of scRNA-seq as in the case of RNA-seq count data [56]. Mahata et al. enumerated many biological insights from scRNA-seq data based on gene-gene correlation [58]. Xue et al. used the WGCNA package to identify many co-expressed modules present in the scRNA-seq data and these have been found to be biologically meaningful [59]. Crow et al. performed a co-expression analysis of scRNA-seq to understand the key drivers of gene-gene connectivity and replicability in co-expression networks [60]. They performed co-expression and meta-analysis of scRNA-seq data, where they compared co-expression networks constructed from individual cell types and all of the cell types, and then assessed the cell-state and variations in functional connectivity. Interestingly, they found that scRNA-seq co-expression shows lower functional connectivity than that derived from RNA-seq. This happens due to batch effects or incomplete transcriptome coverage, inherent to current scRNA-seq protocols [29], [60], [61]. In addition, single-cell co-expression aligns less with known biology due to greater specificity. Methods used to assess scRNA-seq based on co-expression include topological and GO based analysis [60]. In GRN inference of scRNA-seq, pseudo-time information and expression dynamics are generally used [62]. SCODE is a tool based on the transformation of linear ordinary differential equations and linear regression [62]. It performs GRN construction better than competing algorithms such as GENIE3 [63]. LEAP is another algorithm based on pseudo-time ordering of cells to construct co-expression networks from scRNA-seq data [64]. Differential network analysis of scRNA-seq data can reveal many interesting findings, which have not been fully explored so far. SCDE is a tool to perform gene level analysis of scRNA-seq data [65]. It uses a Bayesian method to compare two groups of single cells by taking into account variability and amplification biases present in scRNA-seq data.

4 Co-Expression Analysis

Co-expression analysis finds genes that have a tendency to exhibit similar expressions in a group of samples. It finds a smaller network containing modules from the co-expression networks of genes. It facilitates associating a gene of unknown function with a biological process. Another interesting application of co-expression analysis is prioritization of candidate disease genes or gene groups. Different types of analysis can be performed on a co-expression network. These types are shown in Fig. 2.

Module based analysis is the main goal of co-expression analysis because modules are more informative with respect

Authorized licensed use limited to: Universidad Nacional Autonoma De Mexico (UNAM). Downloaded on September 18,2024 at 20:53:55 UTC from IEEE Xplore. Restrictions apply.

to the specific regulatory mechanism of a biological process [3]. Co-expression analysis can be of two types [3]: (1) Guide gene approach and (2) Non-targeted approach. In the first category, the approach starts with a prior set of genes responsible for a particular biological problem based on the literature and these genes are considered target or guide genes during co-expression analysis. They help find genes which are correlated with the genes of interest. In the second category, no prior knowledge is used to find modules from network based on its topology. The guided gene approach also does not exclude the possibility of a module being a part of a larger and more densely connected module.

Limitations of co-expression analysis include: (1) causality information or any distinction between regulatory and regulated genes cannot be found from gene co-expression networks, and (2) modules which show different co-expression patterns across different tissues or conditions are not detectable in co-expression analysis because the correlation measure used during computation of pairwise correlation of genes may show lower correlation between a pair of genes due to the lack of proper correlation signal in other conditions or tissues for the same pair of genes. These limitations can be overcome by differential co-expression analysis, which is discussed in details in Section 6.

4.1 Types of Co-Expression Analysis

Most commonly used approaches for co-expression analysis are discussed below.

4.1.1 Clustering to Find Modules

Clustering is a process of grouping genes which have similar expression patterns across samples. A cluster often represents a biological process [66] or it may be phenotype specific [67]. Clustering can find tissue or genotype specific clusters which are related to biochemical pathways [68]. A cluster in the context of a co-expression network is shown in Fig. 2a.

Weighted Gene Correlation Network Analysis (WGCNA) is the most widely used package for module finding [69]. It applies hierarchical clustering on a co-expression network to find modules. This tool has applied to RNA-seq/scRNA-seq data and it can effectively find modules which are responsible for biological processes. cMonkey$_2$ is a tool to detect co-regulated gene modules for an organism [70]. This user-friendly tool is automated and systematic. ConGEMs is a tool for exploring biomarker modules from transcriptomic data by detecting co-expression modules using association rule mining and weighted similarity scores [71]. FGMD is based on a hierarchical clustering algorithm [41]. It was developed to capture actual biological observations. It has shown the best performance on different datasets for most functional pathway enrichment tests. This algorithm has been applied to both microarray and RNA-seq datasets, and modules constructed from these two types of datasets have many common genes.

Modules in a network can be characterized by topological properties such as network density [72], degree distribution [73], clustering coefficient [74], betweenness centrality [75], closeness centrality [76], and eigenvector centrality [77]. In topological analysis, identification of hubs, differential features of networks, and differential functional enrichment between two different conditions are performed. Cui et al. present details of these topological properties in [78].

4.1.2 Hub Genes Finding

Hub genes are highly connected genes in a co-expression network. Generally, a module contains many genes. So, it is important to find which gene(s) can represent the behavior of the module effectively. Finding the so-called hub of a set of genes in a module is used for this purpose. Each intra-hub genes are more relevant than other genes from the perspective of the functioning of the network [79]. Intra-hub genes represent a function of a module more effectively and inter-hub genes are central to the entire co-expression network. Samples of intra-and inter-hub genes can be viewed in Fig. 2b. To identify hub genes, measures such as betweenness centrality, degree, and connectivity are often used. Hub gene identification in a co-expression network helps identification of genes responsible for diseases such as cancer [80], type 2 diabetes [81] and other diseases [82]. However, in the case of multiple hub genes, it is difficult to say which hub gene is responsible for a phenotype variation. DHGA can identify hub genes and perform differential hub analysis in a co-expression network to group hub genes based on connectivity [83]. WiPer is a package that can identify hub genes based on connectivity scores [84]. Langfelder et al. performed a study to determine when hub gene selection is better than meta-analysis (aggregating results from separate studies without regards to gene-gene relationships [67]) for two objectives (1) gaining biological insight, and (2) for relevance to basic research [67]. They found that to gain biological insights, intra-hub gene selection is better than meta-analysis with respect to consensus modules. However, meta-analysis is more useful with the consensus network approach. Ballouz et al. performed a comparison by constructing microarray and RNA-seq co-expression networks [45]. They found that the correlation between the degrees of nodes from the two networks is negative and both networks have different hub gens. MRHCA is a tool based on a non-parametric approach that can identify hub genes in a co-expression network and modules associated with each hub. This tool has been is applied to both microarray and RNA-seq data, and it has found satisfactory performance with both.

4.1.3 Guilt by Association

Biological processes are carried out by modules of interacting molecules. The Guilt by association (GBA) approach provide one framework to predict unknown gene function of the genes present in the extracted module from co-expression network by considering the gene functions of interacting genes [85], [86]. Disease genes are usually co-expressed with other disease genes. So, using the guilt by association approach, we can identify potential disease genes. The guilt by association approach can be used to associate function to a poorly annotated gene present in the same co-expressed module. E.g., if some of the genes present in a module are responsible for a disease, the remaining genes can be analyzed to determine whether they are responsible for the disease or not using the GBA approach, as shown in Fig. 2c. When the GBA approach is used one must remember that all genes present in a module are not responsible for a biological function. Usually only $< 20\%$ of the genes present in a module is responsible for its main biological function [86].

A gene co-expression network can be analyzed in terms of how much qualitative functional information can be

Authorized licensed use limited to: Universidad Nacional Autonoma De Mexico (UNAM). Downloaded on September 18,2024 at 20:53:55 UTC from IEEE Xplore. Restrictions apply.

extracted from the network using GBA. In a gene network, only a few interactions and outliers are responsible for the functional information encoded by that network [86]. These cannot be used to study the generalized behavior of the rest of the network elements although they affect the performance of GBA in practice. EGAD is a tool to evaluate functional properties of networks based on GBA [87]. This tool can predict most informative groups of genes from co-expression networks and can assess how well a gene network groups a known set of genes.

4.1.4 Regulatory Network Construction

A Gene Regulatory Network predicts relationships among genes present in gene expression data [88]. In a GRN, nodes represent objects of interest and edges represent relationships such as repression or activation among them. Such a relationship between a pair of genes can happen via the binding of a Transcription Factor (TF) or via a signalling cascade or pathway [88]. A few tools are available for regulatory network construction such as ARACNE [89] and GENIE3 [63]. ARACNE constructs a regulatory network by eliminating indirect connections among genes. GENIE3 constructs a regulatory network with the help of TF expression patterns. Statistical methods used for GRN reconstruction can be found in [90]. More details can be found in Section 5.

4.1.5 Integrative Analysis

There are many other types of data such as protein-protein interaction (PPI) data, methylome data, transcript factors, miRNA, and expression quantitative loci (eQTLs) that affect gene expression and modules in the network [91]. Integration of different data types with a co-expression network can be viewed in Fig. 2e. CoRegNet [92] and cMonkey$_2$ [70] are two tools that allow integration of mentioned data types in co-expression analysis. NICCE is a tool for gene co-expression analysis in citrus that integrates gene-expression networks, functional enrichment analysis and gene expression information to predict gene function [93]. Medina et al. combined co-expression network analysis with functional enrichment analysis to identify potential biomarkers involved in Type 1 diabetes (T1D) [94]. Cui et al. performed a hybrid analysis to detect disease related genes by exploiting topological properties of PPI networks [78].

CoExpNetViz enables researchers to use their own data for cross-species co-expression network analysis [95]. It takes query genes from one or multiple species and gene expression data as input, and gives a file in the output which can be visualized in Cytoscape in network form. wTO is a tool that allows weighted network calculation for both positive and negative correlation [96]. It also allows computation of consensus networks, which enables integration of networks derived from different studies and provides a visualization facility.

4.1.6 Preservation and Disruption Analysis

After module identification in co-expression network analysis, module preservation and disruption studies can reveal hidden information present in the co-expression modules across conditions. Preservation and disruption analysis of a module is shown in Fig. 2f. In this figure, the module in the first condition is highly co-expressed while in the second condition, the co-expression strengths have vanished. Some widely used statistics for this purpose are reported in Table 1 [97]. Comparison of co-expression networks of multiple species such as mice and human has revealed conservation and divergence at disease, tissue and pathway levels. So, conservation is a powerful approach in co-expression network analysis and it has been found that modules in co-expressed networks are conserved by natural selection [98]. Such studies can be used to identify potential genes and processes responsible for phenotypic variations [99]. Stuart et al. performed co-expression analysis and found that a large number of co-expression relationships are conserved across evolution in human, fly, worm and yeast, and many of these relationships are in turn responsible for many core biological functions [100]. ORTom is a conservation-based tool which can be used to prioritize candidate genes to targeted studies and to infer functional association among genes [101]. More details on module preservation be found in [102].

Co-expression Network Toolkit (CoNekT), an open source, user-friendly web server can perform comparison analysis of co-expression networks across species [103]. It identifies modules responsible for evolution from one species to another. ComPlEx is a tool to perform conservation and divergence analysis of co-expression networks across species to understand underlying species differentiation [104].

5 Network Inference to Reconstruct Transcriptional Networks

Network Inference (NI) is also an interesting application of gene expression data analysis. It reconstructs transcriptional networks, i.e., GRNs through reverse engineering. It also facilitates integration of other data types such as PPI, miRNA, TF, and DNA methylation data based on prior knowledge, such as predicted or known relationships or knowledge sources [108]. GRNs help in finding relationships among genes and identifying which genes are activators or repressors of gene expressions of other genes. ARACNE [89] and GENIE3 [63] are two widely used tools for GRN construction. As mentioned earlier scRNA-seq data do not have temporal information. So, capturing dynamic features of underlying cellular mechanisms is difficult. But, for reverse engineering to reconstruct transcriptional networks, it is possible to find pseudo temporal observations using the cell time ordering algorithm [109]. Moignard et al. used state transition graphs and diffusion maps to construct GRNs [110]. Chen et al. developed an stability based network component analysis (sNCA) approach to identify regulatory networks and this approach is quite robust to false positives in finding GRNs [111]. CRNET is another tool for discovering GRNs in large-scale ChIP-seq and time-course RNA-seq data [112]. mAPC-GibbsOS is another GRN identification method [113]. It uses clustering to modularize regulatory structures in the first step, and refinement in the next step to find true target genes. SCENIC is a tool for GRN construction from scRNA-seq data that can also help identify stable cell states [114]. scRNA-seq based GRNs can help discover more interesting biological insights, which may not be possible by RNA-seq based GRNs [56]. Two TFs may independently regulate expressions of a set of genes; the expression at one but non-expression at the other cannot be detected in RNA-seq based GRNs because they will be considered co-expressed. But, in scRNA-seq this type of scenario is easily distinguishable [56]. SpidermiR is a package that allows users to query and download GRNs and it offers

Authorized licensed use limited to: Universidad Nacional Autonoma De Mexico (UNAM). Downloaded on September 18,2024 at 20:53:55 UTC from IEEE Xplore. Restrictions apply.

TABLE 1
Statistics for Module Preservation and Disruption Measurement

Name	Description
Density	It measures compactness of a module. Density is defined as the average connection strength of all pairs of genes present in a module.
Clustering coefficient	Clustering co-efficient of a node quantifies the interconnectedness of its neighboring nodes. The average of clustering coefficients of all nodes measures the modularity of a module.
Maximum adjacency ratio	It generally deals with a weighted network and can be used to check whether a hub gene is connected to a large number of genes with moderate connection strengths or a few genes with strong connection strengths [105]. By computing the average MAR, we can distinguish among the connectivity patterns. More the correlation between MAR of nodes in two modules across conditions, then the preservation is higher.
Sign preserving mean correlation	This module preservation strategy takes element-wise multiplication of the correlation matrices for the same module in two different samples groups and then takes the mean. Higher mean means more preservation.
Eigengene based measure	Eigengene is an arbitrary gene represents a module very well. To calculate preservation, one can take the correlation between the eigengenes of the corresponding modules of different phenotypes. Higher correlation means more preservation.
Intra-modular connectivity	The sum of connection strengths of neighbors of a node. To investigate preservation, correlation of intra-modular connectivity of nodes present in the same module of different phenotypes can be used. More correlation implies higher preservation.
Fitting coefficient (R^2) [69]	Scale-free topology criteria of a network can be quantified using the fitting coefficient of linear regression.
Global network connectivity [106]	This characteristic path length is the average of the shortest paths between all pairs of nodes in a network.
Zsummary [107]	It quantifies the amount of preservation of a module of reference in a test network by considering density and connectivity of that module in both networks.
MedianRank [107]	It gives the ranking of a reference network modules in a test network. The most preserved module gets the top rank.

easy work with both GRNs and miRNAs, and integrates the information about differentially expressed genes from the Cancer Genomic Atlas [115].

6 Differential Co-Expression Analysis

Understanding differential co-expression of gene expression data underlying molecular mechanisms can be valuable. Such analysis can help discover disease mechanisms and underlying regulatory dynamics, which cannot be discovered using differential expression analysis alone [116]. It also helps in understanding disease progression, disease subclassification, and identification potential biomarkers responsible for phenotypic variation or critical diseases. This approach identifies a group of genes which show varying co-expression patterns over different conditions such as development stages [59], tissue types [117] and disease states [118]. A graphical representation of differential co-expression is presented in Fig. 3. Disease related genes have tissue specific expressions. So, this analysis can help identify genes that show co-expression in both similar and different types of tissues [117]. The genes which have different co-expression patterns in different sample groups, are the regulatory genes responsible for phenotypic variations observed in different conditions. Many network-based studies have been performed to identify unique disease or tissue type specific co-expression networks using differential co-expression analysis. A few applications of differential co-expression analysis include disease sub-classification to obtain prognostic markers and analysis of datasets in which sub-populations are unknown, e.g., large-scale single-cell RNA-seq data analysis. This approach is sensitive to noise.

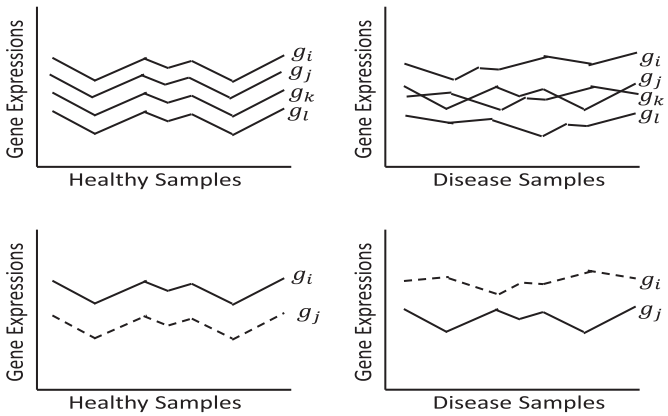

Fig. 3. Differential co-expression types: (a) A gene cluster is differentially co-expressed across samples groups, and (b) a pair of co-expressed genes is differentially co-expressed under different sample groups.

We feel that there are scopes for exploring appropriate multi-objective co-expression cluster formation strategy to find diferentially co-expressed modules. Moreover, the weighted combination of semantic and statistical proximity measures in finding association between a pair of genes also promising. The possible multi-objectives are high modularity at control condition, higher change in correlation across conditions and considered genes are differentially expressed. After finding such modules, one can perform processing like functional enrichment analysis, which includes GO enrichment analysis, pathway analysis, and exploration of relevant wet lab facts available in literature.

Generally, to perform differential co-expression analysis, condition specific co-expression networks are constructed, and this requires different decision and threshold settings for the networks related to healthy and disease states,

Authorized licensed use limited to: Universidad Nacional Autonoma De Mexico (UNAM). Downloaded on September 18,2024 at 20:53:55 UTC from IEEE Xplore. Restrictions apply.

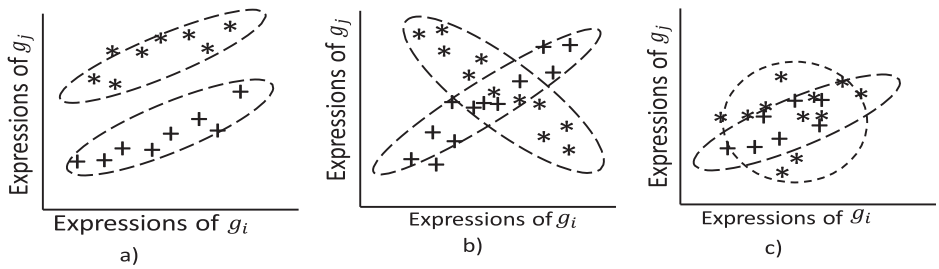

Fig. 4. Different ways of differential co-expression for a pair of genes [123]. Figure a): Shift, b): Inverse/*Cross*, and c): Re-wiring.

respectively. This makes differential co-expression analysis complicated. So, instead of establishing co-expression pattern changes across conditions, one can directly perform differential co-expression analysis by testing whether changes in co-expression across conditions are significant or not [119]. Childs et al. performed condition dependent and condition independent co-expression network analysis, and concluded that condition dependent co-expression network modules are more useful for biological interpretation [120]. Some methods do not construct a network to assess differential co-expressions of genes. GSNCA can be used for such analysis, but sample groups need to defined a priori [121]. If no knowledge is provided about sample groups, biclustering can be used to identify modules by classifying the samples into groups. EPIG-Seq is a tool to identify biologically meaningful co-expressed genes in RNA-seq data across biological conditions, i.e., experimental groups [122].

6.1 Differential Co-Expression Patterns and Detection Strategies

Different differential co-expression patterns include [123]: (1) Shifting-and/or-scaling, where the correlation between a pair of genes is similar across conditions but expression values have clear changes through shifting-and/or-scaling; (2) Cross/Inverse, where the genes are positively correlated in one condition and negatively correlated in another condition; and (3) Re-wiring, where the genes show high correlation in one condition but the correlation vanishes in another condition. These can be viewed in Fig. 4. Genes with significant differential co-expression are used to understand phenotype differences and may play role as biomarkers in studying regulatory impacts of genes on other genes across conditions. Small differences in expression patterns cannot be detected using simple differential expression analysis methods such as *t*-test or ANOVA, and so, differential expression analysis is not always sufficient to find biomarkers [124]. This a fact that has emerged in differential co-expression in recent years.

Differential co-expression analysis can be divided into two major types depending on whether (a) a pair of co-expressed genes is differentially expressed [123], [125], or (b) a gene cluster is differentially co-expressed [123]. These two types are shown in Fig. 3. In the first case, a gene pair may follow shifting-and/or-scaling, cross/inverse or re-wiring relationships in differential co-expression analysis. In the second case, a group of genes may be co-expressed in one sample group and the co-expression may vanish in another sample group. The different ways in which differential co-expression of a group of genes can occur are presented in Fig. 5. To find that a pair of genes is differentially co-expressed, first we have to quantify the connection strength between that pair of genes using an appropriate method, and then based on the changes in connection strength across conditions, we can predict whether that pair is differentially co-expressed or not. GSAR is a tool for gene set analysis to detect shift, scale, or cross/inverse correlation structures [126]. It also provides visualization facility to examine the changes in the correlation structure of a gene set between two conditions. It also highlights influential hub genes.

Detection techniques for differential co-expression can be categorized into two: targeted and untargeted. In the targeted approach, no module finding is performed. In this approach, correlation changes across conditions of a predefined gene set are quantified, whereas, in the untargeted approach, a group of co-expressed modules are first identified and then correlation changes are quantified. This approach can identify novel differentially co-expressed modules. Some untargeted approaches are EBcoexpress [127], ROS-DET [128], DiffCoEx [129], and DEICER [118]. These methods do not consider the possibility of shift patterns, which can be addressed by BFDCA [130].

The presence of cross/inverse pattern is the most commonly present differential co-expression scenario. Some widely used methods to detect cross/inverse relationships for a pair of genes are as follows.

1) *Changes in correlation coefficient across conditions (CCC):* To detect that a pair of genes is differentially co-expressed, the changes in the correlation coefficients are quantified across conditions. This is a two step procedure: (1) computing correlation between the pair of genes under each condition, and (2) taking the differences between the computed correlations. These two steps can be formulated as given below for a pair of genes (g_i, g_j) under normal (x) and disease (y) conditions. A larger the value of $CCC(g_i, g_j)$ speaks more in favor of then pair of genes being differentially co-expressed

$$CCC(g_i, g_j) = |CC_x(g_i, g_j) - CC_y(g_i, g_j)|. \quad (1)$$

To calculate the correlation coefficient, one can use Pearson or Spearman correlation, or bi-weight midcorrelation.

2) *Changes in entropy across conditions (CEC):* Randomness in a system is measured using entropy [131]. Entropy is proportional to randomness. Like CCC, we can quantify change in entropy across conditions, for a pair of gene (g_i, g_j) under normal (x) and disease (y) conditions. Larger values of $CEC(g_i, g_j)$ indicate more in favor of the possibility that the pair of genes is differentially co-expressed

$$CEC(g_i, g_j) = \frac{\sum_m I_m}{m} - I, \quad (2)$$

Authorized licensed use limited to: Universidad Nacional Autonoma De Mexico (UNAM). Downloaded on September 18,2024 at 20:53:55 UTC from IEEE Xplore. Restrictions apply.

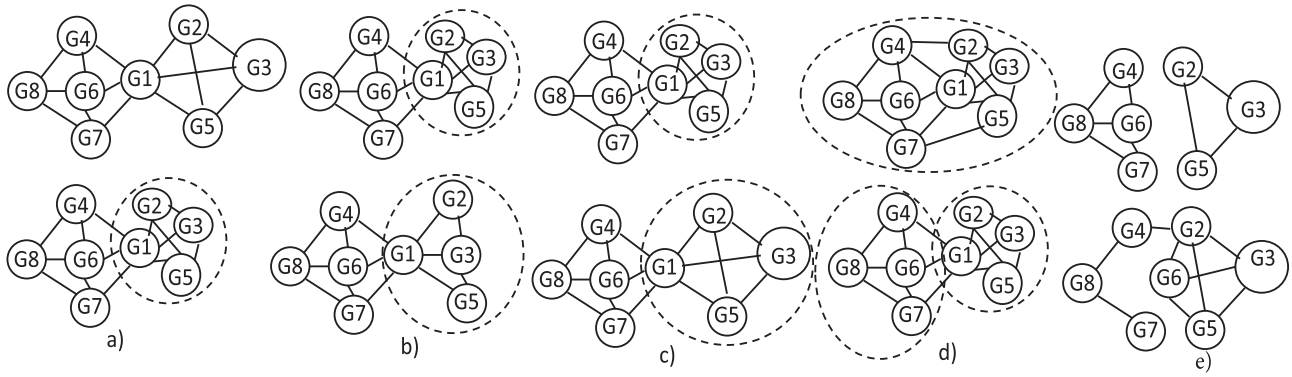

Fig. 5. Differential co-expression can occur in the ways shown above. The first row corresponds to one group of samples in normal/control condition, and the next row corresponds to second group of samples in disease condition. a) A module (group of co-expressed genes) is present in only one of the sample groups such as samples related to a disease state. b) Change in the structure of the modules with respect to conditions such as normal and disease conditions. c) Correlation strength changes between the genes present in a module where different states show different correlation strengths. d) One larger module present in a normal samples group is split in another group of samples. e) Some genes present in a module in normal state change their interaction partners with respect to the change in the sample group.

where k is the number of conditions and I_m and I are the Shannon's entropy values that can be formulated as below:

$$I_m = -\frac{1+|CC_m|}{2} log\left(\frac{1+|CC_m|}{2}\right) + \frac{1-|CC_m|}{2} log\left(\frac{1-|CC_m|}{2}\right) \quad (3)$$

$$I = -\frac{1+|CC|}{2} log\left(\frac{1+|CC|}{2}\right) + \frac{1-|CC|}{2} log\left(\frac{1-|CC|}{2}\right). \quad (4)$$

Here, m is the set of conditions and CC_m is the correlation coefficient between a pair of genes (g_i, g_j) at condition m. CC is the correlation coefficient between a pair of genes (g_i, g_j) considering all conditions.

3) *Expected conditional F-statistic*: F-statistic can be used to quantify the degree of differential expression of a gene. It is used to compare the ratios of variances within groups (K) to variances between groups. Later, this same test can also be applied to find differentially co-expressed genes by considering conditional expectations [132]. λ can be defined using the F-statistic for gene g as below:

$$\lambda = \frac{K-1}{N-K} F = \frac{\sum_k N_k (\bar{g_k} - \bar{g})^2}{\sum_{j,k} N_k (g_{jk} - \bar{g_k})^2}. \quad (5)$$

λ is for an individual gene. To apply it for a pair of genes (g_i, g_j), we must replace the mean and the variance with the conditional mean and variance of a gene g_i given $g_j = g_j'$. To quantify the score of a pair of genes (g_i, g_j) that are differentially co-expressed, the formulation is as follows:

$$S_{ECF(g_i,g_j)} = E_{g_j}(\lambda_{g_i|g_j=g_j'}) + E_{g_i}(\lambda_{g_j|g_i=g_i'}). \quad (6)$$

4) *Interaction test:* Differential interactions between a pair of genes (g_i, g_j) can be assessed by an interaction test [133].

DC of Multiple Genes. We can consider more than two genes during differential co-expression checking. It is possible that a group of genes is highly correlated in one condition and correlation vanishes in another condition. A score is assigned to the group of genes, indicating whether that gene set is differentially co-expressed or not. Several methods are discussed below.

1) *Change in correlation co-efficient:* Correlation co-efficients between all possible pairs of genes in the group are calculated under different conditions and then differences in correlations are quantified. If the differences are high, we can say that the group of genes is differentially co-expressed.

   Watson et al. developed a tool named CoXpress, based on Pearson correlation co-efficient and t-statistic [134]. It quantifies the correlation co-efficient and t-statistic differences among all pairs of genes across conditions to make a determination about differential co-expressions among multiple genes.
2) *Mutual information change*: Correlation co-efficient may not be suitable to check whether a group of genes is differentially expressed or not because for a randomly expressed pair of genes, the correlation co-efficient may give values close to zero. Mutual information measures randomness of dependency between two variables. More the mutual information, the higher the dependency between two variables. So, instead of correlation co-efficient, we can use mutual information to check whether a group of genes is differentially co-expressed or not.
3) *Ratio of root mean square errors*: RMSE can detect differentially expressed modules by checking how well a model fits the data. If the data has a co-expression pattern, it will fit better than data with random patterns. If the data fits the model well, RMSE will be lower. So, by comparing RMSE across conditions we can detect differentially co-expressed genes.

Mitsunori et al. performed a review of 9 different methods for both differential co-expression between a pair of genes and multiple genes in a group [123]. They reported that for pairwise detection of differential co-expression between genes, ROS-DET [128] was the most powerful

Authorized licensed use limited to: Universidad Nacional Autonoma De Mexico (UNAM). Downloaded on September 18,2024 at 20:53:55 UTC from IEEE Xplore. Restrictions apply.

method. This method is based on bi-weight midcorrelation and can work well even with random and range biased data [123]. For multiple genes, they reported that CoXpress [134] and GSCA [135] were the best methods [123]. When they used Spearman correlation in CoXpress, results were most robust.

Generally, diverse metrics have been used to detect differential co-expression. There are many differential co-expression techniques, and these are usually metric specific. A single correlation measure which can work well in all scenarios remains elusive. Gonzalez et al. performed a comparative study of 6 metrics used to detect differential co-expression between a pair of genes, and reported that metrics differed in detection performance and computational time [136]. They used 3 real time cancer datasets and found no metric that performed well consistently across all datasets. They provide an R implementation named DIFferential COexpression NETworks (DifCoNet) that has an easy-to-use function for differential correlation analysis. They reported that any metric that does not filter correlation information performed better. The metrics they used include difference in the number of correlation [137], Kolmogorov Smirnov distance [138], sum of large correlations differences, euclidean distance, Kullback Leibler divergence [139], and Adjacency difference [129].

6.2 Types of Differential Co-Expression Analysis

Differential co-expression analysis can be performed in two ways based on whether sample groups are classified or not.

1) *Generalized Singular Value Decomposition (GSVD):* This type of differential co-expression analysis identifies modules of genes based on spectral decomposition. In this technique, first, a summary is prepared for all samples and genes to represent expression data using a small number of variables. Next, the significance of representative variables is compared between two datasets. If no significant difference is found, they are co-expressed, else they are differentially co-expressed. More recently, Higher Order (HO)-GSVD has been used to compare between more than two datasets [140].
2) *Differential co-expression without grouping:* When knowledge of sample groups is not provided, this type of differential co-expression analysis technique is used to find differentially expressed biclusters. This approach does not require prior sample classification. So, it is generally superior to co-expression analysis methods [141]. Effectiveness of a biclustering technique generally depends on the datasets, and so, it is difficult to choose an appropriate biclustering technique for this approach. Gao et al. applied biclustering without correcting confounders in gene expression datasets and found condition-specific co-expression networks which were differentially expressed across conditions [142]. DeBi is a biclustering algorithm that can discover differentially expressed biclusters in which genes are co-expressed in a subset of samples and the co-expression pattern is lost in the rest of the samples [143].

6.3 Tools for (Differential) Co-Expression Analysis

The most widely used programs for differential clustering analysis are WGCNA [144], DICER [118], DiffCoEx [129] and DINGO [145]. These programs first identify co-expressed modules considering all samples and then these modules are compared with a defined set of samples such as samples representing disease states or tissue types. E.g., if a pair of genes show high positive correlation in normal condition and negative correlation in a disease condition, then there is a high chance that these genes are responsible for the disease transformation. Siska et al. developed a tool named Discordant and reported that Spearman's correlation is the best method for differential correlation analysis in RNA-seq data [146].

Most differential co-expression methods perform analysis on the full space in samples of each sample group. These approaches may miss patterns that only cover a subset of samples, i.e., a sub-space in one or more sample groups. Fang et al. performed sub-space based differential co-expression analysis and proposed a method based on association analysis framework to find the sub-space of differential co-expression patterns [147]. They also found that discovered sub-space patterns are functionally enriched in terms of pathways and are related to microRNA and TFs. Many tools for (differential) co-expression analysis have been reported and most were initially developed for microarrays and later adapted for RNA-seq data analysis. We present 19 of them in Table 2 with description and sources of availability.

Differential expression analysis and co-expression analysis are applied extensively in scRNA-seq data and now differential co-expression analysis needs attention from researches. Differential expression analysis and differential co-expression analysis identify genes and genes modules, respectively showing varying expression patterns across conditions. Some differential expression analysis tools such as DESeq2 [4] and EdgeR [148] developed for bulk RNA-seq give very good results in scRNA-seq data [149]. Xue et al. identified interesting modules present in scRNA-seq using the WGCNA [144] package [59]. There are some possibilities for adapting tools for (differential) co-expression analysis developed for bulk RNA-seq in scRNA-seq data by considering the fact that multiple cell types are included in one scRNA-seq dataset. MAST is a tool especially developed for scRNA-seq data for differential expression analysis of genes and gene modules [150]. DiffCoEx [129], CEMiTool [151] and BFDCA [130] use WGCNA to identify modules from differential networks, and these two tools can be adapted to RNA-seq/scRNA-seq datasets with two sample groups. SeqMade [152], especially developed for RNA-seq data, can be adapted for scRNA-seq datasets of two types of sample groups. The superiority of this tool was established by comparing with tools developed for microarray data analysis. This tool can also be adapted to scRNA-seq datasets, provided there are well-defined cell groups. CoXpress [134], GSCNA [121], GSVD [153] and GSCA [135] are targeted approaches and they use a variety of statistics to test whether a given gene set is differentially expressed or not in another sample group. These tools can also be adapted to scRNA-seq data by considering the fact that multiple cell types are included in one scRNA-seq dataset. CoXpress can work with for more than two sample groups. DiffCoMO [154] is multi-objective and MultiDCoX [155] is multi-factor based differential co-expression tools. DiffCoMO was tested on microarray datasets with more than 2 sample groups, and it can be applied to scRNA-seq data, given cells group information. MultiDCoX can be used to scRNA-seq by considering cell types as co-factors, i.e., similar samples will be treated as one sample group.

Authorized licensed use limited to: Universidad Nacional Autonoma De Mexico (UNAM). Downloaded on September 18,2024 at 20:53:55 UTC from IEEE Xplore. Restrictions apply.

TABLE 2
(Differential) Co-Expression Analysis Tools

Tools	Years	Description	Availability
CoXpress [134]	2006	It finds all co-expressed modules in one sample group and can test whether these modules are co-expressed in another sample group or not.	http://coxpress.sf.net
WGCNA [144]	2008	It uses Pearson correlation measure (as default) to find the correlation between a pair of genes. It is the most widely used method, and it builds a weighted co-expression network, and ensures scale free topology. Botia et al. [156] provide refinement in a module finding step of WGCNA and present a tool. Their refinement step substantially improves the module properties in terms of finding accurate modules and functional analysis.	https://cran.r-project.org/web/packages/WGCNA/index.html
GSCA [135]	2009	Gene Set Co-expression Analysis requires a predefined group of genes to check whether they are differentially expressed or not. It also computes the false discovery rates of interesting gene groups. This approach highly depends on the correlation measure used.	www.biostat.wisc.edu/~kendzior/GSCA/
DiffCoEx [129]	2010	It finds modules of genes which have different partners in different sample groups using WGCNA with little modification.	Refer supplementary of [129].
SDCA [147]	2010	Only a subset of samples are considered from each condition in differential co-expression analysis. It uses biclustering in the background to identify biclusters present in each condition and later analyzes the changes in expression of biclusters across samples groups.	http://vk.cs.umn.edu/SDC/
HO-GSVD [140]	2011	It is the same as GSVG. The only difference is that it can be used across multiple sample groups. It outperforms WGCNA and DiffCoEx [157].	http://www.wolfram.com/products/player/
GSCNA [121]	2013	It can test whether a predefined set of genes are expressed differently or not in two sample groups.	Available upon request
DICER [118]	2013	It can form modules which show correlation differently in different sample groups. A larger interconnected module in one sample group is divided into multiple smaller modules in another sample group. It performs better than DiffCoEx and CoXpress [118].	http://acgt.cs.tau.ac.il/dicer/
DCGL v2.0 [158]	2013	This tool is suitable for both microarray and RNA-seq data and can be used to identify novel regulators from gene expression data. It facilitates both differential regulation and differential co-expression analysis.	http://cran.r-project.org/web/packages/DCGL/index.html
DINGO [145]	2015	It groups genes that are expressed differently in a predefined subset of samples, e.g., samples related to the development states of a disease.	odin.mdacc.tmc.edu/vbaladan
Biclustering [159]	2015	A group of methods that identify modules that are unique to a subpopulation of samples without the need for prior grouping of samples.	
CoGA [160]	2015	Co-expression Graph Analyzer finds the spectral distributions of gene co-expression graphs using entropy to quantify the randomness of a graph structure, which helps discriminate among the classes of graphs. In addition, it allows structural property based comparison of co-expression networks. It also provides a visualization facility for networks and has lower false positive rate.	http://cran.r-project.org/
DGCA [161]	2016	Differential Gene Correlation Analysis measures and analyzes differential correlations of gene pairs across conditions. It helps in identification of novel pathways and causal genes in complex biological systems and diseases.	https://github.com/andymckenzie/DGCA
BFDCA [130]	2017	It is based on the Bayes Factor approach. It can take care of different differential co-expression patterns such as Shift, Cross/Inverse and Re-wiring using one Bayes factor. It has high accuracy and robustness in detecting differential co-expression patterns and can outperform existing methods.	http://dx.doi.org/10.17632/jdz4vtvnm3.1
MultiDCoX [155]	2017	This is the only multi-factor differential co-expression analysis approach. It is a time efficient approach and can quantify the influence of differential co-expression.	R function will be available upon request.
Discordant [146]	2017	It is based on Expectation Maximization and a mixture model to find the differential correlation in count data. Authors of this tool report that Spearman's correlation is the best correlation measure for count data. Sub-sampling extension to the Expectation Maximization algorithm has limitations such as the number of independent pairs is limited by the dimension of the data. Though the ranks can be similar, higher posterior probability is shown by sub-samples.	http://bioconductor.org/packages/discordant
SeqMade [152]	2017	It is based on Generalized Linear Model (GLM) for differential co-expression of RNA-seq count data.	https://cran.r-project.org/web/packages/SeqMADE/
DiffCoMO [154]	2017	It is a novel multi-objective framework to capture differentially co-expressed modules across different conditions, i.e., different stages of HIV-1. The objective function emphasizes the distance between co-expression patterns across the stages.	—
CEMiTool [151]	2018	This fast and easy-to-use tool can be used to discover and analyze co-expression modules. This tool was applied to 8 RNA-seq datasets, and it outperformed widely used tools such as WGCNA, DiffCoEx, DICER and DINGO. In addition, it has the ability to perform pathway analysis.	https://doi.org/10.18129/B9.bioc.CEMiTool

6.4 Differential Networking

In a differential network, a connection represents the amount of change in correlation strength in a pair of genes across conditions with the provision that the change is significant. The main aim of differential network analysis is to find a sub-network (i.e., a subset of connected genes) that operates abnormally in each disease state. In a differential network as shown in Fig. 6, only those edges which are present in either healthy or disease network are retained. For biomarker identification, the assumptions are that a gene which changes a significant number of neighbors across conditions is more likely to be an interesting biomarker [1]. So, the most straightforward biomarker identification technique is to check the degrees of the nodes [162]. DIfferential Network Analysis (DINA) is able to identify a group of genes whose co-expression pattern is condition specific [163]. It can also predict which TF is responsible for pathway condition specific co-expression. FDvDe can detect

Authorized licensed use limited to: Universidad Nacional Autonoma De Mexico (UNAM). Downloaded on September 18,2024 at 20:53:55 UTC from IEEE Xplore. Restrictions apply.

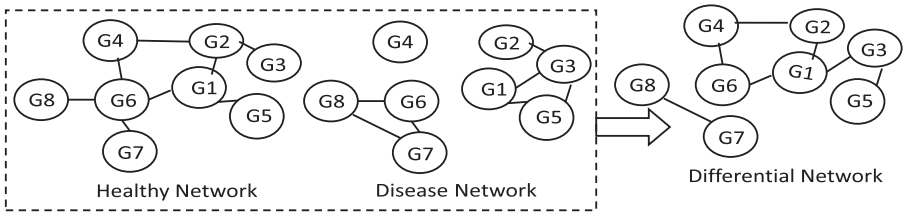

Fig. 6. Differential network. In this network, only those edges among genes are retained which are present either healthy or disease network.

differential co-expression by constructing differential networks from a condition specific co-expression network by aggregating differential nodes and differential edges [116]. Then it applies GO Term Hierarchical Clusters (GTHC) to identify functional modules based on semantic similarity, and follows by topological analysis to identify hub genes and hub pathways, which are interesting for disease mechanisms. CoDiNA can compare any number of networks related to different stages of a disease [164]. It can distinguish between links present in all networks, links specific to a network and links that change their sign. DiffRank can rank genes based on their importance to make two networks distinguishable in terms of differential connectivity and differential centrality [165]. Some state-of-the-art tools for differential network analysis include Differential Co-Expressed Links (DCELs) [166], Differential Co-Expression Network (DCEN) [166], DCglob [167] (DeteCtion of global change of the network topology), and DCloc (DeteCtion of the local changes of the network) [167].

6.5 Differential Expression versus Differential Co-Expression versus Differential Networking versus Differential Connectivity

In differential co-expression analysis, the correlation between a pair of genes changes across conditions, and it may happen in the absence of differential expression [168]. It means that differential expression analysis may not be able to detect changes in regulatory patterns of genes across conditions. This situation generally arises when the sample groups are related to the same tissue types [169]. Differential expression analysis detects a large number of genes showing significant altered expressions across conditions [50]. On other hand, differential co-expression analysis identifies a set of modules which are differentially expressed across conditions. It can also be used to prioritize specific phenotype related genes. Differential network analysis helps find a sub-network that has high changes in correlation strengths for a pair of genes across conditions. Differential co-expression analysis can help find the underlying regulatory dynamics, which cannot be discovered by differential analysis [116]. Differential connectivity analysis finds how many interaction partner genes change across conditions. It helps identify genes that are involved in different phenotype variations. It has been found that genes which show high differential connectivity across conditions do not show up among the top significant differentially expressed genes [170]. Kakati et al. reported that genes which are in the periphery of a co-expression network are more interesting and may be causal genes [171]. These genes also have lower connectivity, and most are differentially expressed. Differential connectivity analysis can reveal genes responsible for a trait, which are not revealed by differential expression analysis. For example, loss of connectivity in a disease network with respect to the normal ones is a trait in different kinds of cancer [137]. A graphical representation of differential connectivity can be viewed in Fig. 7. Oldham et al. performed a study in the conservation and evolution of different regions of brains of human and chimpanzee based on gene co-expression networks [172]. They found that differential expression and differential connectivity show significant correlation, and they identified key drivers of evolutionary change using differential connectivity analysis. They also claimed that differential connectivity analysis is significantly superior to differential expression analysis to find evolutionary changes, i.e., key drivers responsible for evolution.

7 Hybrid Analysis

In differential expression analysis, individual genes are considered, but interactions among genes are ignored whereas in general, biomolecules do not act alone. This limitation is addressed in (differential) co-expression analysis very well. These two approaches may extract both dependent and independent information [173]. We can deploy them either separately or combine them based on the dependencies in the information extracted and can get new opportunities for dissecting complex disease mechanisms. DECODE is an approach to combine differential and differential co-expression analysis [173]. It categorizes genes into four groups with either high or low differential expression and differential co-expression. It gives some biologically meaningful results, which may not be possible to find by differential expression or differential co-expression analysis alone. INDEED is another hybrid analysis tool which combines differential expression and differential network analysis to build a sparse differential network for biomarker identification [174]. It gives more reproducible results across independent datasets.

One limitation of hybrid analysis is how to choose the appropriate threshold for selection of highly differentially expressed genes and highly differentially co-expressed pairs of genes. Anglani et al. found that combined analysis using both differential connectivity and differential expression analysis improves the classical pathway analysis to obtain biological insights [137]. Yu et al. presented the computational workflow to identify differentially co-expressed modules using differential expression analysis

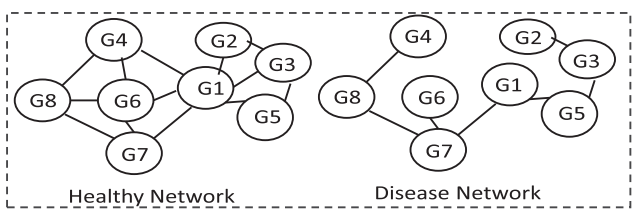

Fig. 7. Differential connectivity. A few genes (e.g., G1 and G6) lose their connectivity with many neighbors when they move to the disease network.

Authorized licensed use limited to: Universidad Nacional Autonoma De Mexico (UNAM). Downloaded on September 18,2024 at 20:53:55 UTC from IEEE Xplore. Restrictions apply.

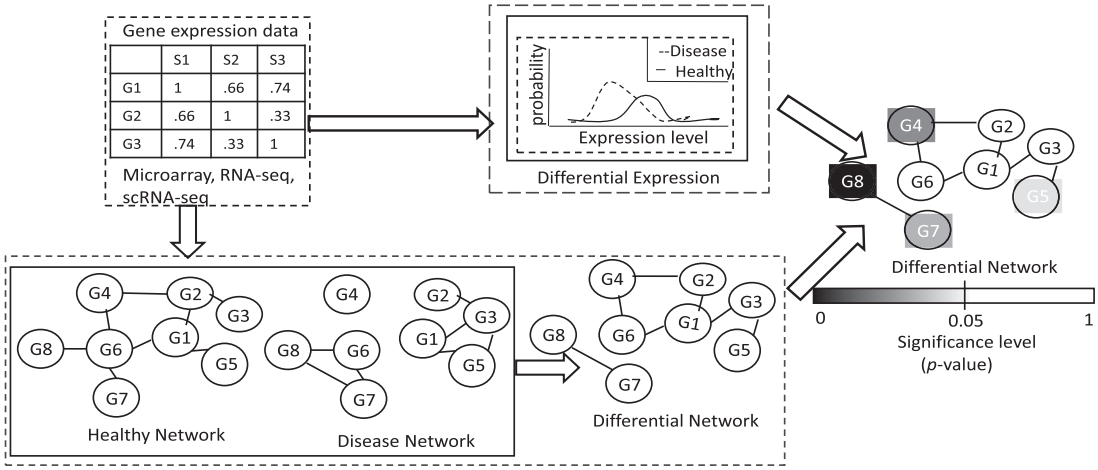

Fig. 8. Hybrid analysis. Here, differential expression and differential network analysis are performed on gene expression data from healthy and disease conditions, and then genes are considered based on the significance level i.e., p-value ≤ 0.05 to get biological insights.

and SVD across eight cancer types, and found that four modules are functionally enriched [175]. Fang et al. combined differential and differential co-expression analysis to identify functionally related optimal gene sets in cervical cancer [176]. They also found that the combination made their method superior to individual differential and differential co-expression analysis. Drag et al. performed both differential and co-expression analysis on RNA-seq data of pig to identify boar taint (BT) related biomarkers [177]. A graphical representation of hybrid analysis is presented in Fig. 8.

8 Biological Interpretation and Functional Analysis

Functional annotation analysis extracts biological information from a set of genes that have been identified through unsupervised techniques. Generally, Gene Ontology (GO) and pathway-based analyses are widely used in functional analysis to interpret the outcome of gene expression analysis. GO provides a shared understanding of a domain of information. Pathway and ontology-based analyses consist of steps such as mapping, statistical testing, and presentation [178]. Researchers have developed many analysis tools for each step and these include ArrayX, GOFish [179], GOMIner [180], FatiGO [181], GOTree Machine [182], GOAL [183], FuncAssociate [184], MAPPFinder [185], DAVID [178], and GFINDer [186].

GO-Module [187] helps interpret GO terms by reducing the number of GO terms up to 4 times lower than the number of input GO terms. It provides an organized output and keeps only those GO terms which contain truly positive results. EnrichNet is a web application, that addresses limitations of traditional functional enrichment analysis such as difficulty in dealing with overlapping genes and proteins and disregarding genes with missing annotation, discarding network structures of physical interactions and tissues specific to gene and protein sets [188].

Generally, groups of genes share biological processes. To understand these processes, knowledge of metabolic pathways and signalling pathways is a very powerful resource. Metabolic pathways are enumerated in KEGG [189] whereas signalling pathways are enumerated in BioCarta [190].

Pathway and network analyses help in getting functional insights from a biological system. Pathways and networks have certain conceptual distinctions. In a pathway, interactions are biochemical reactions and events related to regulation and signals, whereas network interactions are genome-wide interactions derived from datasets [191]. These approaches have a number of benefits in relation to the analysis of genomics data [191]. (1) They aggregate all related genes as modules, which may decrease the need of more statistical tests. (2) They help in interpretation of results in more depth and more easily. (3) Biomarker identification becomes easier. (4) Results from related datasets become comparable more meaningfully. (5) Integration of multiple data types is possible during analysis. Major approaches for pathway and network analysis include [191]: (1) fixed-gene set enrichment analysis, (2) de novo network construction and clustering and (3) network-based modelling.

In the past years, to improve functional understanding and discovery of meaningful information from large-scale genomic data through biological semantics, different types of data such as biomedical annotation, external repositories and clinical data have been integrated.

9 Guidelines for the Practitioner

Following our extensive review of tools and relevant literature, the following sections provide some informal guidelines and recommendations for gene expression data analysis.

9.1 Preprocessing

a) Proper understanding of datasets is essential for gene ID conversion, verification and elimination of samples with missing metadata, to support effective downstream analysis.
b) Outliers and genes with very low read counts need to be removed before performing downstream analysis.
c) Selection of techniques for proper normalization, imputation and transformation of data can help find more biological insights.
d) No imputation method performs well for all types of scRNA-seq data [192]. So, it may be useful to combine the results of multiple algorithms to impute data.

Authorized licensed use limited to: Universidad Nacional Autonoma De Mexico (UNAM). Downloaded on September 18,2024 at 20:53:55 UTC from IEEE Xplore. Restrictions apply.

e) We recommend that a practitioner should use scRNA-seq specific normalization methods and update the scRNA-seq analysis pipeline to take full advantage of advances in RNA-seq technology [23].

9.2 Gene Expression Data Analysis

a) Higher quality of samples and sizes are necessary to identify tissue-specific or disease-specific modules and regulators and to increase functional connectivity analysis of the co-expression networks [45].
b) More than 10 million reads per sample are ideal for co-expression analysis [45].
c) It has been suggested that one should consider at least 20 samples to perform co-expression analysis of RNA-seq [45].
d) ROS-DET [128] is the most powerful method for pairwise detection of differential co-expression of genes.
e) In reality, only $< 20\%$ genes present in the modules are responsible for main biological functions [86]. So, the recommendation is to use the guilt by association approach more carefully so that we do not predict the function of a gene wrongly.
f) If one assesses the quality of gene networks using the guilt by association approach, checking the effect of critical edges in the network is essential [86].
g) Functional clustering may be more useful for identification of (differentially) co-expressed genes.
h) Integrated analysis of differential expression and (differential) co-expression can reveal more interesting genes which may not be found by individual analysis [137].
i) It is a good idea to use a consensus of differential expression analysis tools to select differentially expressed genes before (differential) co-expression analysis.
j) Integrated data analysis helps in better understanding of co-expression patterns responsible for a disease and the regulatory mechanisms.

9.3 Establishment of Results

a) Analysis of findings which are obtained using benchmark datasets can be carried out to ascribe more biological meanings to the results.
b) Theoretical and clinical establishment of results based on theoretical evidence and clinical knowledge is highly desirable.
c) Validation of results with the support of possible evidence found in high quality research is essential to establish analysis findings.

10 Research Issues and Challenges

Based on our extensive reading in preparation for writing this survey and our extensive experience in biological data analysis [50], [171], [193], [194], [195], [196], we observe that in the past decade, many methods and tools have been developed to support effective (differential) co-expression analysis. However, there are some important issues and research challenges that still need attention. In this section, we briefly highlight some such issues and challenges.

a) The presence of an abundant number of 0s, i.e., zero inflation in RNA-seq data is an issue when calculating correlation between a pair of genes.
b) Choosing optimal sample size and read depth required for co-expression analysis of RNA-seq is challenge. A low number of replicates in RNA-seq data is problematic for effective differential co-expression analysis.
c) An efficient and powerful imputation method which is robust, scalable and applicable to all types of data and situations is currently unavailable [192].
d) A robust and effective normalization method for scRNA-seq data is needed. It is necessary to a identify a best method for scRNA-seq data normalization [23].
e) Performance of a normalization method is dataset specific [16]. So, there is a need for an effective normalization method for RNA-seq/scRNA-seq data. It will also be ideal to have a technique, which can automatically identify a suitable method from a a set of normalization methods for a particular RNA-seq/scRNA-seq data.
f) Different normalization methods have different impacts on co-expression analysis of RNA-seq. So, selection of an appropriate normalization method is important.
g) Correlation measures developed for microarrays are not suitable in RNA-seq because of the discrete data distribution of RNA-seq data. Though logarithmic and variance stabilizing transformations can be used to address such problems, there is a chance of losing information. So, there is a need for a robust correlation measure which can work well with RNA-seq data without transformation.
h) A single correlation measure which can work well in all situations for differential co-expression patterns is missing [136].
i) In inferring undirected networks, the selection of a proper co-expression measure and module finding method is an an issue [90] because they greatly influence downstream analysis.
j) Co-expression modules are not perfectly modular because it has been found that many members of a module may have a strong correlation with members of other modules [197]. So, there is a need for effective methods to address any perturbation effects from outside of a module.
k) Generally, many genes in identified modules from co-expression networks do not have annotation information, creating difficulties in inferring association of such genes with traits [33].
l) Co-expression analysis of splice variants is challenging if multiple splice variants are associated with the same exon because we cannot predict which splice variant is expressed leading to the biases in results.
m) Topological changes in co-expression networks need more attention so that one can understand cell development and obtain biological insights for identification of potential biomarkers.
n) The performance of differential co-expression analysis techniques depends on the datasets, species types, disease states, and so, optimal technique selection is time consuming and challenging, and there is need for a differential co-expression techniques with

Authorized licensed use limited to: Universidad Nacional Autonoma De Mexico (UNAM). Downloaded on September 18,2024 at 20:53:55 UTC from IEEE Xplore. Restrictions apply.

generic features that support investigation of a large number of critical diseases.

o) A combination of microarray and RNA-seq data may increase biological knowledge [40]. So, there is a need for attention from the research community for a proper framework to combine analyses to find more hidden biological knowledge.

p) Statistical evaluation of pathway and network methods in patient care needs further investigation. Proper methods for communicating genomic information to clinicians and patients remain unclear.

q) A framework to support integrative analysis that combines both transcriptomic and clinical data for a family of critical diseases needs to be explored.

11 Conclusions

In this paper, we have presented an overview of best practices in the analysis of (differential) co-expression, co-expression networks, differential networking, and differential connectivity that can be discovered in microarrays and RNA-seq data, and shed some light on the analysis of scRNA-seq data as well. We have discussed co-expression analysis for RNA-seq along with a comparison of analysis of co-expression networks in microarrays versus RNA-seq. We discussed differential co-expression and differential networking along with a comparison of differential expression, differential networking and differential connectivity. Throughout the paper, we included many tools used for analyzing microarry, RNA-seq, and scRNA-seq data. Reverse engineering to reconstruct transcriptional network is also discussed. Preprocessing and co-expression analysis of scRNA-seq data along with useful tools and applications are also included. We discussed biological interpretation and functional analysis to extract biological information from a set of given genes that are identified through analysis. We presented some recommendations, and guidelines for the analyst. Finally, we outlined several research issues and challenges for future researchers and practitioners.

References

[1] A. Reverter, A. Ingham, S. A. Lehnert, et al., "Simultaneous identification of differential gene expression and connectivity in inflammation, adipogenesis and cancer," *Bioinf.*, vol. 22, no. 19, pp. 2396–2404, 2006.

[2] A. van der Graaf, L. Franke, U. Võsa, S. van Dam, and J. P. de Magalhães, "Gene co-expression analysis for functional classification and gene–disease predictions," *Briefings Bioinf.*, vol. 19, no. 4, pp. 575–592, Jan. 2017, doi: 10.1093/bib/bbw139.

[3] K. Aoki, Y. Ogata, and D. Shibata, "Approaches for extracting practical information from gene co-expression networks in plant biology," *Plant Cell Physiology*, vol. 48, no. 3, pp. 381–390, 2007.

[4] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biol.*, vol. 15, no. 12, 2014, Art. no. 550.

[5] T. Aittokallio, "Dealing with missing values in large-scale studies: Microarray data imputation and beyond," *Briefings Bioinf.*, vol. 11, no. 2, pp. 253–264, 2009.

[6] L. Zhu, J. Lei, B. Devlin, K. Roeder, et al., "A unified statistical framework for single cell and bulk RNA sequencing data," *The Annals Applied Statist.*, Institute of Mathematical Statistics, vol. 12, no. 1, pp. 609–632, 2018.

[7] D. van Dijk, J. Nainys, R. Sharma, et al., "MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data," BioRxiv, 2017, Art. no. 111591.

[8] W. Gong, I.-Y. Kwak, P. Pota, N. Koyano-Nakagawa, and D. J. Garry, "DrImpute: Imputing dropout events in single cell RNA sequencing data," *BMC Bioinf.*, vol. 19, no, 1, p. 220, 2018.

[9] M. Huang, J. Wang, E. Torre, et al., "Gene expression recovery for single cell RNA sequencing," BioRxiv, 2017, Art. no. 138677.

[10] S. Prabhakaran, E. Azizi, A. Carr, et al., "Dirichlet process mixture model for correcting technical variation in single-cell gene expression data," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1070–1079.

[11] L. Zhang and S. Zhang, "Comparison of computational methods for imputing single-cell RNA-sequencing data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 2018.

[12] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical Bayes methods," *Biostatist.*, vol. 8, no. 1, pp. 118–127, 2007.

[13] M. J. Nueda, A. Ferrer, and A. Conesa, "ARSyN: A method for the identification and removal of systematic noise in multifactorial time course microarray experiments," *Biostatist.*, vol. 13, no. 3, pp. 553–566, 2012.

[14] A. Conesa, P. Madrigal, S. Tarazona, et al., "A survey of best practices for RNA-seq data analysis," *Genome Biol.*, vol. 17, no. 1, 2016, Art. no. 13.

[15] C. Filloux, M. Cédric, P. Romain, et al., "An integrative method to normalize RNA-Seq data," *BMC Bioinf.*, vol. 15, no. 1, 2014, Art. no. 188.

[16] T. Eder, F. Grebien, and T. Rattei, "NVT: A fast and simple tool for the assessment of RNA-seq normalization strategies," *Bioinf.*, vol. 32, no. 23, pp. 3682–3684, 2016.

[17] J. Zyprych-Walczak, A. Szabelska, L. Handschuh, et al., "The impact of normalization methods on RNA-Seq data analysis," *BioMed Res. Int.*, vol. 2015, 2015, Art. no. 621690.

[18] A. Mortazavi, B. A. Williams, K. McCue, et al., "Mapping and quantifying mammalian transcriptomes by RNA-Seq," *Nature Methods*, vol. 5, no. 7, 2008, Art. no. 621.

[19] B. Li, V. Ruotti, R. M. Stewart, et al., "RNA-Seq gene expression estimation with read mapping uncertainty," *Bioinf.*, vol. 26, no. 4, pp. 493–500, 2009.

[20] J. H. Bullard, E. Purdom, K. D. Hansen, et al., "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments," *BMC Bioinf.*, vol. 11, no. 1, 2010, Art. no. 94.

[21] M. D. Robinson and A. Oshlack, "A scaling normalization method for differential expression analysis of RNA-seq data," *Genome Biol.*, vol. 11, no. 3, 2010, Art. no. R25.

[22] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome Biol.*, vol. 11, no. 10, 2010, Art. no. R106.

[23] C. A. Vallejos, D. Risso, A. Scialdone, et al., "Normalizing single-cell RNA sequencing data: Challenges and opportunities," *Nature Methods*, vol. 14, no. 6, 2017, Art. no. 565.

[24] B. Ding, L. Zheng, Y. Zhu, et al., "Normalization and noise reduction for single cell RNA-seq experiments," *Bioinf.*, vol. 31, no. 13, pp. 2225–2227, 2015.

[25] S. Katayama, V. Töhönen, S. Linnarsson, et al., "SAMstrt: Statistical test for differential expression in single-cell transcriptome with spike-in normalization," *Bioinf.*, vol. 29, no. 22, pp. 2943–2945, 2013.

[26] R. Bacher, L.-F. Chu, N. Leng, et al., "SCnorm: A quantile-regression based approach for robust normalization of single-cell RNA-seq data," BioRxiv, 2016, Art. no. 090167.

[27] I. Zwiener, B. Frisch, and H. Binder, "Transforming RNA-Seq data to improve the performance of prognostic gene signatures," *PloS One*, vol. 9, no. 1, 2014, Art. no. e85150.

[28] V. M. Kvam, P. Liu, and Y. Si, "A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data," *Amer. J. Botany*, vol. 99, no. 2, pp. 248–256, 2012.

[29] E. Pierson and C. Yau, "ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis," *Genome Biol.*, vol. 16, no. 1, 2015, Art. no. 241.

[30] V. Y. Kiselev, K. Kirschner, M. T. Schaub, et al., "SC3: Consensus clustering of single-cell RNA-seq data," *Nature Methods*, vol. 14, no. 5, 2017, Art. no. 483.

[31] S. L. Carter, C. M. Brechbühler, M. Griffin, et al., "Gene co-expression network topology provides a framework for molecular characterization of cellular state," *Bioinf.*, vol. 20, no. 14, pp. 2242–2250, 2004.

[32] Z.-P. Liu, Y. Wang, X.-S. Zhang, et al., "Network-based analysis of complex diseases," *IET Syst. Biol.*, vol. 6, no. 1, pp. 22–33, 2012.

[33] M. Filteau, S. A. Pavey, J. St-Cyr, et al., "Gene coexpression networks reveal key drivers of phenotypic divergence in lake whitefish," *Mol. Biol. Evol.*, vol. 30, no. 6, pp. 1384–1396, 2013.

Authorized licensed use limited to: Universidad Nacional Autonoma De Mexico (UNAM). Downloaded on September 18,2024 at 20:53:55 UTC from IEEE Xplore. Restrictions apply.

[34] S. Kumari, J. Nie, H.-S. Chen, et al., "Evaluation of gene association methods for coexpression network construction and biological knowledge discovery," *PloS One*, vol. 7, no. 11, 2012, Art. no. e50411.

[35] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C.-F. Chen, "A new method to measure the semantic similarity of GO terms," *Bioinf.*, vol. 23, no. 10, pp. 1274–1281, 2007.

[36] D. J. Allocco, I. S. Kohane, and A. J. Butte, "Quantifying the relationship between co-expression, co-regulation and gene function," *BMC Bioinf.*, vol. 5, no. 1, 2004, Art. no. 18.

[37] J. K. Mandal, S. Mukhopadhyay, and P. Dutta, *Multi-Objective Optimization: Evolutionary to Hybrid Framework*. Berlin, Germany: Springer, 2018.

[38] H. Wang, F. Azuaje, O. Bodenreider, et al., "Gene expression correlation and gene ontology-based similarity: An assessment of quantitative relationships," in *Proc. IEEE Symp. Comput. Intell. Bioinf. Comput. Biol.*, 2004, pp. 25–31.

[39] S. Freytag, J. Gagnon-Bartsch, T. P. Speed, et al., "Systematic noise degrades gene co-expression signals but can be corrected," *BMC Bioinf.*, vol. 16, no. 1, 2015, Art. no. 309.

[40] L. López-Kleine and C. González-Prieto, "Challenges analyzing RNA-seq gene expression data," *Open J. Statist.*, vol. 6, no. 04, pp. 628–636, 2016.

[41] D. Jin and H. Lee, "FGMD: A novel approach for functional gene module detection in cancer," *PloS One*, vol. 12, no. 12, 2017, Art. no. e0188900.

[42] O. D. Iancu, S. Kawane, D. Bottomly, et al., "Utilizing RNA-Seq data for de novo coexpression network inference," *Bioinf.*, vol. 28, no. 12, pp. 1592–1597, 2012.

[43] F. M. Giorgi, C. Del Fabbro, and F. Licausi, "Comparative study of RNA-seq-and microarray-derived coexpression networks in Arabidopsis thaliana," *Bioinf.*, vol. 29, no. 6, pp. 717–724, 2013.

[44] R. S. Sekhon, R. Briskine, C. N. Hirsch, et al., "Maize gene atlas developed by RNA sequencing and comparative evaluation of transcriptomes based on RNA sequencing and microarrays," *PLoS One*, vol. 8, no. 4, 2013, Art. no. e61005.

[45] S. Ballouz, W. Verleyen, and J. Gillis, "Guidance for RNA-seq co-expression network construction and analysis: Safety in numbers," *Bioinf.*, vol. 31, no. 13, pp. 2123–2130, 2015.

[46] S. van Dam, T. Craig, and J. P. de Magalhães, "GeneFriends: A human RNA-seq-based gene and transcript co-expression database," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D1124–D1132, 2014.

[47] H. K. Yalamanchili, Z. Li, P. Wang, M. P. Wong, J. Yao, and J. Wang, "SpliceNet: Recovering splicing isoform-specific differential gene networks from RNA-Seq data of normal and diseased samples," *Nucleic Acids Res.*, vol. 42, no. 15, pp. e121–e121, 2014.

[48] L. Folkersen, T. Kyriakou, A. Goel, J. Peden, A. Mälarstig, G. Paulsson-Berne, A. Hamsten, A. Franco-Cereceda, A. Gabrielsen, P. Eriksson, et al., "Relationship between CAD risk genotype in the chromosome 9p21 locus and gene expression. Identification of eight new ANRIL splice variants," *PloS One*, vol. 4, no. 11, 2009, Art. no. e7677.

[49] M. Brüss, M. Barann, M. Hayer-Zillgen, T. Eucker, M. Göthert, and H. Bönisch, "Modified 5-HT3A receptor function by co-expression of alternatively spliced human 5-HT3A receptor isoforms," *Naunyn-Schmiedeberg's Archives Pharmacology*, vol. 362, no. 4/5, pp. 392–401, 2000.

[50] H. A. Chowdhury, D. K. Bhattacharyya, and J. K. Kalita, "Differential expression analysis of RNA-seq reads: Overview, taxonomy and tools," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, to be published, doi: 10.1109/TCBB.2018.2873010.

[51] O. D. Iancu, A. Colville, D. Oberbeck, et al., "Cosplicing network analysis of mammalian brain RNA-Seq data utilizing WGCNA and Mantel correlations," *Frontiers Genetics*, vol. 6, 2015, Art. no. 174.

[52] C. Dai, W. Li, J. Liu, et al., "Integrating many co-splicing networks to reconstruct splicing regulatory modules," *BMC Syst. Biol.*, vol. 6, no. 1, 2012, Art. no. S17.

[53] J. Petereit, S. Smith, F. C. Harris, et al., "Petal: Co-expression network modelling in R," *BMC Syst. Biol.*, vol. 10, no. 2, 2016, Art. no. 51.

[54] S. Proost, A. Krawczyk, and M. Mutwil, "LSTrAP: Efficiently combining RNA sequencing data into co-expression networks," *BMC Bioinf.*, vol. 18, no. 1, 2017, Art. no. 444.

[55] M. E. Smoot, K. Ono, J. Ruscheinski, et al., "Cytoscape 2.8: New features for data integration and network visualization," *Bioinf.*, vol. 27, no. 3, pp. 431–432, 2010.

[56] O. Stegle, S. A. Teichmann, and J. C. Marioni, "Computational and analytical challenges in single-cell transcriptomics," *Nature Rev. Genetics*, vol. 16, no. 3, pp. 133–145, 2015.

[57] D. Grün, A. Lyubimova, L. Kester, et al., "Single-cell messenger RNA sequencing reveals rare intestinal cell types," *Nature*, vol. 525, no. 7568, 2015, Art. no. 251.

[58] B. Mahata, X. Zhang, A. A. Kolodziejczyk, et al., "Single-cell RNA sequencing reveals T helper cells synthesizing steroids de novo to contribute to immune homeostasis," *Cell Rep.*, vol. 7, no. 4, pp. 1130–1142, 2014.

[59] Z. Xue, K. Huang, C. Cai, et al., "Genetic programs in human and mouse early embryos revealed by single-cell RNA [thinsp] sequencing," *Nature*, vol. 500, no. 7464, pp. 593–597, 2013.

[60] M. Crow, A. Paul, S. Ballouz, et al., "Exploiting single-cell expression to characterize co-expression replicability," *Genome Biol.*, vol. 17, no. 1, 2016, Art. no. 101.

[61] D. Grün, L. Kester, and A. Van Oudenaarden, "Validation of noise models for single-cell transcriptomics," *Nature Methods*, vol. 11, no. 6, 2014, Art. no. 637.

[62] H. Matsumoto, H. Kiryu, C. Furusawa, et al., "SCODE: An efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation," *Bioinf.*, vol. 33, no. 15, pp. 2314–2321, 2017.

[63] A. Irrthum, L. Wehenkel, P. Geurts, et al., "Inferring regulatory networks from expression data using tree-based methods," *PLoS One*, vol. 5, no. 9, 2010, Art. no. e12776.

[64] A. T. Specht and J. Li, "LEAP: Constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering," *Bioinf.*, vol. 33, no. 5, pp. 764–766, 2016.

[65] P. V. Kharchenko, L. Silberstein, and D. T. Scadden, "Bayesian approach to single-cell differential expression analysis," *Nature Methods*, vol. 11, no. 7, 2014, Art. no. 740.

[66] Y. Chen, J. Zhu, P. Y. Lum, et al., "Variations in DNA elucidate molecular networks that cause disease," *Nature*, vol. 452, no. 7186, pp. 429–435, 2008.

[67] P. Langfelder, P. S. Mischel, and S. Horvath, "When is hub gene selection better than standard meta-analysis?" *PloS One*, vol. 8, no. 4, 2013, Art. no. e61505.

[68] A. Fukushima, M. Kusano, H. Redestig, et al., "Metabolomic correlation-network modules in Arabidopsis based on a graph-clustering approach," *BMC Syst. Biol.*, vol. 5, no. 1, 2011, Art. no. 1.

[69] B. Zhang, S. Horvath, et al., "A general framework for weighted gene co-expression network analysis," *Statistical Appl. Genetics Mol. Biol.*, vol. 4, no. 1, 2005, Art. no. 1128.

[70] D. J. Reiss, C. L. Plaisier, W.-J. Wu, et al., "cMonkey2: Automated, systematic, integrated detection of co-regulated gene modules for any organism," *Nucleic Acids Res.*, vol. 43, no. 13, pp. e87–e87, 2015.

[71] S. Mallik and Z. Zhao, "ConGEMs: Condensed gene co-expression module discovery through rule-based clustering and its application to carcinogenesis," *Genes*, vol. 9, no. 1, 2017, Art. no. 7.

[72] J. C. Mitchell, *Social Networks in Urban Situations: Analyses of Personal Relationships in Central African Towns*. Manchester, U.K.: Manchester Univ. Press, 1969.

[73] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Sci.*, vol. 286, no. 5439, pp. 509–512, 1999.

[74] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, 1998, Art. no. 440.

[75] U. Brandes, "A faster algorithm for betweenness centrality," *J. Math. Sociology*, vol. 25, no. 2, pp. 163–177, 2001.

[76] L. C. Freeman, S. P. Borgatti, and D. R. White, "Centrality in valued graphs: A measure of betweenness based on network flow," *Social Netw.*, vol. 13, no. 2, pp. 141–154, 1991.

[77] P. Bonacich, "Power and centrality: A family of measures," *Amer. J. Sociology*, vol. 92, no. 5, pp. 1170–1182, 1987.

[78] Y. Cui, M. Cai, Y. Dai, et al., "A hybrid network-based method for the detection of disease-related genes," *Physica A: Statistical Mech. Appl.*, vol. 492, pp. 389–394, 2018.

[79] R. Albert, H. Jeong, and A.-L. Barabási, "Error and attack tolerance of complex networks," *Nature*, vol. 406, no. 6794, pp. 378–382, 2000.

[80] E.-Y. Oh, S. M. Christensen, S. Ghanta, et al., "Extensive rewiring of epithelial-stromal co-expression networks in breast cancer," *Genome Biol.*, vol. 16, no. 1, 2015, Art. no. 128.

[81] M. P. Keller, Y. Choi, P. Wang, et al., "A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility," *Genome Res.*, vol. 18, no. 5, pp. 706–716, 2008.

Authorized licensed use limited to: Universidad Nacional Autonoma De Mexico (UNAM). Downloaded on September 18,2024 at 20:53:55 UTC from IEEE Xplore. Restrictions apply.

[82] I. Voineagu, X. Wang, P. Johnston, et al., "Transcriptomic analysis of autistic brain reveals convergent molecular pathology," *Nature*, vol. 474, no. 7351, pp. 380–384, 2011.
[83] S. Das, P. K. Meher, A. Rai, et al., "Statistical approaches for gene selection, hub gene identification and module interaction in gene co-expression network analysis: An application to aluminum stress in soybean (Glycine max L.)," *PloS One*, vol. 12, no. 1, 2017, Art. no. e0169605.
[84] F. J. Azuaje, "Selecting biologically informative genes in co-expression networks with a centrality score," *Biol. Direct*, vol. 9, no. 1, 2014, Art. no. 12.
[85] C. J. Wolfe, I. S. Kohane, and A. J. Butte, "Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks," *BMC Bioinf.*, vol. 6, no. 1, 2005, Art. no. 227.
[86] J. Gillis and P. Pavlidis, "Guilt by association is the exception rather than the rule in gene networks," *PLoS Comput. Biol.*, vol. 8, no. 3, 2012, Art. no. e1002444.
[87] S. Ballouz, M. Weber, P. Pavlidis, et al., "EGAD: Ultra-fast functional analysis of gene networks," *Bioinf.*, vol. 33, no. 4, pp. 612–614, 2016.
[88] J. Linde, S. Schulze, S. G. Henkel, et al., "Data-and knowledge-based modeling of gene regulatory networks: An update," *EXCLI J.*, vol. 14, 2015, Art. no. 346.
[89] A. A. Margolin, I. Nemenman, K. Basso, et al., "ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context," *BMC Bioinf.*, vol. 7, no. 1, 2006, Art. no. S7.
[90] Y. R. Wang and H. Huang, "Review on statistical methods for gene network reconstruction using expression data," *J. Theoretical Biol.*, vol. 362, pp. 53–61, 2014.
[91] K. Glass, C. Huttenhower, J. Quackenbush, et al., "Passing messages between biological networks to refine predicted interactions," *PloS One*, vol. 8, no. 5, 2013, Art. no. e64832.
[92] R. Nicolle, F. Radvanyi, and M. Elati, "CoRegNet: Reconstruction and integrated analysis of co-regulatory networks," *Bioinf.*, vol. 31, no. 18, pp. 3066–3068, 2015.
[93] D. C. Wong, C. Sweetman, and C. M. Ford, "Annotation of gene function in citrus using gene expression information and co-expression networks," *BMC Plant Biol.*, vol. 14, no. 1, 2014, Art. no. 186.
[94] I. R. Medina and Z. Lubovac-Pilav, "Gene co-expression network analysis for identifying modules and functionally enriched pathways in type 1 diabetes," *PloS One*, vol. 11, no. 6, 2016, Art. no. e0156006.
[95] O. Tzfadia, T. Diels, S. De Meyer, et al., "CoExpNetViz: Comparative co-expression networks construction and visualization tool," *Frontiers Plant Sci.*, vol. 6, 2016, Art. no. 1194.
[96] D. M. Gysi, A. Voigt, T. de Miranda Fragoso, E. Almaas, and K. Nowick, "wTO: An R package for computing weighted topological overlap and consensus networks with an integrated visualization tool," *BMC Bioinf.*, vol. 19, no. 1, p. 392, 2018.
[97] S. Ray and S. Bandyopadhyay, "Discovering condition specific topological pattern changes in coexpression network: An application to HIV-1 progression," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 13, no. 6, pp. 1086–1099, Nov./Dec. 2016.
[98] G. A. Singer, A. T. Lloyd, L. B. Huminiecki, et al., "Clusters of co-expressed genes in mammalian genomes are conserved by natural selection," *Mol. Biol. Evol.*, vol. 22, no. 3, pp. 767–775, 2004.
[99] G. Monaco, S. van Dam, J. L. C. N. Ribeiro, et al., "A comparison of human and mouse gene co-expression networks reveals conservation and divergence at the tissue, pathway and disease levels," *BMC Evol. Biol.*, vol. 15, no. 1, 2015, Art. no. 259.
[100] J. M. Stuart, E. Segal, D. Koller, et al., "A gene-coexpression network for global discovery of conserved genetic modules," *Sci.*, vol. 302, no. 5643, pp. 249–255, 2003.
[101] L. Miozzi, P. Provero, and G. P. Accotto, "ORTom: A multi-species approach based on conserved co-expression to identify putative functional relationships among genes in tomato," *Plant Mol. Biol.*, vol. 73, no. 4/5, pp. 519–532, 2010.
[102] B. Li, Y. Zhang, Y. Yu, et al., "Quantitative assessment of gene expression network module-validation methods," *Sci. Rep.*, vol. 5, 2015, Art. no. 15258.
[103] S. Proost and M. Mutwil, "CoNekT: An open-source framework for comparative genomic and transcriptomic network analyses," *Nucleic Acids Res.*, vol. 46, no. W1, pp. W133–W140, May 2018, doi: 10.1093/nar/gky336.
[104] S. Netotea, D. Sundell, N. R. Street, et al., "ComPlEx: Conservation and divergence of co-expression networks in A. Thaliana, Populus and O. Sativa," *BMC Genomics*, vol. 15, no. 1, 2014, Art. no. 106.
[105] S. Horvath and J. Dong, "Geometric interpretation of gene coexpression network analysis," *PLoS Comput. Biol.*, vol. 4, no. 8, 2008, Art. no. e1000117.
[106] I. W. Taylor, R. Linding, D. Warde-Farley, et al., "Dynamic modularity in protein interaction networks predicts breast cancer outcome," *Nature Biotechnol.*, vol. 27, no. 2, pp. 199–204, 2009.
[107] P. Langfelder, R. Luo, et al., "Is my network module preserved and reproducible?" *PLoS Comput. Biol.*, vol. 7, no. 1, 2011, Art. no. e1001057.
[108] D. Gomez-Cabrero, I. Abugessaisa, D. Maier, et al., "Data integration in the era of omics: Current and future challenges," *BMC Syst. Biol.*, vol. 8, no. 2, 2014, Art. no. I1.
[109] A. Ocone, L. Haghverdi, N. S. Mueller, et al., "Reconstructing gene regulatory dynamics from high-dimensional single-cell snapshot data," *Bioinf.*, vol. 31, no. 12, pp. i89–i96, 2015.
[110] V. Moignard, S. Woodhouse, L. Haghverdi, et al., "Decoding the regulatory network of early blood development from single-cell gene expression measurements," *Nature Biotechnol.*, vol. 33, no. 3, pp. 269–276, 2015.
[111] X. Chen, J. Xuan, C. Wang, et al., "Reconstruction of transcriptional regulatory networks by stability-based network component analysis," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 10, no. 6, pp. 1347–1358, Nov./Dec. 2013.
[112] X. Chen, J. Gu, X. Wang, et al., "CRNET: An efficient sampling approach to infer functional regulatory networks by integrating large-scale ChIP-seq and time-course RNA-seq data," *Bioinf.*, vol. 34, no. 10, pp. 1733–1740, 2017.
[113] X. Shi, J. Gu, X. Chen, A. Shajahan, L. Hilakivi-Clarke, R. Clarke, and J. Xuan, "mAPC-GibbsOS: An integrated approach for robust identification of gene regulatory networks," *BMC Syst. Biol.*, vol. 7, no. 5, 2013, Art. no. S4.
[114] S. Aibar, C. B. González-Blas, T. Moerman, H. Imrichova, G. Hulselmans, F. Rambow, J.-C. Marine, P. Geurts, J. Aerts, J. van den Oord, et al., "SCENIC: Single-cell regulatory network inference and clustering," *Nature Methods*, vol. 14, no. 11, p. 1083, 2017.
[115] C. Cava, A. Colaprico, G. Bertoli, et al., "SpidermiR: An R/Bioconductor package for integrative analysis with miRNA data," *Int. J. Mol. Sci.*, vol. 18, no. 2, 2017, Art. no. 274.
[116] S. Meng, G. Liu, L. Su, et al., "Functional clusters analysis and research based on differential coexpression networks," *Biotechnol. Biotechnological Equipment*, vol. 32, no. 1, pp. 171–182, 2018.
[117] E. Pierson, D. Koller, A. Battle, et al., "Sharing and specificity of co-expression networks across 35 human tissues," *PLoS Comput. Biol.*, vol. 11, no. 5, 2015, Art. no. e1004220.
[118] D. Amar, H. Safer, and R. Shamir, "Dissection of regulatory networks that are altered in disease via differential co-expression," *PLoS Comput. Biol.*, vol. 9, no. 3, 2013, Art. no. e1002955.
[119] A. de la Fuente, "From differential expression to differential networking–identification of dysfunctional regulatory networks in diseases," *Trends Genetics*, vol. 26, no. 7, pp. 326–333, 2010.
[120] K. L. Childs, R. M. Davidson, and C. R. Buell, "Gene coexpression network analysis as a source of functional annotation for rice genes," *PloS One*, vol. 6, no. 7, 2011, Art. no. e22196.
[121] Y. Rahmatallah, F. Emmert-Streib, and G. Glazko, "Gene sets net correlations analysis (GSNCA): A multivariate differential coexpression test for gene sets," *Bioinf.*, vol. 30, no. 3, pp. 360–368, 2013.
[122] J. Li and P. R. Bushel, "EPIG-Seq: Extracting patterns and identifying co-expressed genes from RNA-Seq data," *BMC Genomics*, vol. 17, no. 1, 2016, Art. no. 255.
[123] M. Kayano, M. Shiga, and H. Mamitsuka, "Detecting differentially coexpressed genes from labeled expression data: A brief review," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 11, no. 1, pp. 154–167, Jan./Feb. 2014.
[124] M. Zampieri, N. Soranzo, and C. Altafini, "Discerning static and causal interactions in genome-wide reverse engineering problems," *Bioinf.*, vol. 24, no. 13, pp. 1510–1515, 2008.
[125] S. B. Cho, J. Kim, and J. H. Kim, "Identifying set-wise differential co-expression in gene expression microarray data," *BMC Bioinf.*, vol. 10, no. 1, 2009, Art. no. 109.

Authorized licensed use limited to: Universidad Nacional Autonoma De Mexico (UNAM). Downloaded on September 18,2024 at 20:53:55 UTC from IEEE Xplore. Restrictions apply.

[126] Y. Rahmatallah, B. Zybailov, F. Emmert-Streib, et al., "GSAR: Bioconductor package for gene set analysis in R," *BMC Bioinf.*, vol. 18, no. 1, 2017, Art. no. 61.
[127] J. A. Dawson, S. Ye, and C. Kendziorski, "R/EBcoexpress: An empirical Bayesian framework for discovering differential co-expression," *Bioinf.*, vol. 28, no. 14, pp. 1939–1940, 2012.
[128] M. Kayano, I. Takigawa, M. Shiga, et al., "ROS-DET: Robust detector of switching mechanisms in gene expression," *Nucleic Acids Res.*, vol. 39, no. 11, pp. e74–e74, 2011.
[129] B. M. Tesson, R. Breitling, and R. C. Jansen, "DiffCoEx: A simple and sensitive method to find differentially coexpressed gene modules," *BMC Bioinf.*, vol. 11, no. 1, 2010, Art. no. 497.
[130] D. Wang, J. Wang, Y. Jiang, et al., "BFDCA: A comprehensive tool of using Bayes factor for differential co-expression analysis," *J. Mol. Biol.*, vol. 429, no. 3, pp. 446–453, 2017.
[131] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Comput. Commun. Rev.*, vol. 5, no. 1, pp. 3–55, 2001.
[132] Y. Lai, B. Wu, L. Chen, et al., "A statistical method for identifying differential gene–gene co-expression patterns," *Bioinf.*, vol. 20, no. 17, pp. 3146–3155, 2004.
[133] M. Kayano, I. Takigawa, M. Shiga, et al., "Efficiently finding genome-wide three-way gene interactions from transcript-and genotype-data," *Bioinf.*, vol. 25, no. 21, pp. 2735–2743, 2009.
[134] M. Watson, "CoXpress: Differential co-expression in gene expression data," *BMC Bioinf.*, vol. 7, no. 1, 2006, Art. no. 509.
[135] Y. Choi and C. Kendziorski, "Statistical methods for gene set co-expression analysis," *Bioinf.*, vol. 25, no. 21, pp. 2780–2786, 2009.
[136] E.-E. Gonzalez-Valbuena and V. Treviño, "Metrics to estimate differential co-expression networks," *BioData Mining*, vol. 10, no. 1, 2017, Art. no. 32.
[137] R. Anglani, T. M. Creanza, V. C. Liuzzi, et al., "Loss of connectivity in cancer co-expression networks," *PloS One*, vol. 9, no. 1, 2014, Art. no. e87075.
[138] G. K. Kanji, *100 Statistical Tests*. Newbury Park, CA, USA: Sage, 2006.
[139] J. Kasturi, R. Acharya, and M. Ramanathan, "An information theoretic approach for analyzing temporal patterns of gene expression," *Bioinf.*, vol. 19, no. 4, pp. 449–458, 2003.
[140] S. P. Ponnapalli, M. A. Saunders, C. F. Van Loan, et al., "A higher-order generalized singular value decomposition for comparison of global mRNA expression from multiple organisms," *PloS One*, vol. 6, no. 12, 2011, Art. no. e28072.
[141] A. Fiannaca, M. La Rosa, L. La Paglia, et al., "Analysis of miRNA expression profiles in breast cancer using biclustering," *BMC Bioinf.*, vol. 16, no. 4, 2015, Art. no. S7.
[142] C. Gao, I. C. McDowell, S. Zhao, et al., "Context specific and differential gene co-expression networks via Bayesian biclustering," *PLoS Comput. Biol.*, vol. 12, no. 7, 2016, Art. no. e1004791.
[143] A. Serin and M. Vingron, "DeBi: Discovering differentially expressed biclusters using a frequent itemset approach," *Algorithms Mol. Biol.*, vol. 6, no. 1, 2011, Art. no. 18.
[144] P. Langfelder and S. Horvath, "WGCNA: An R package for weighted correlation network analysis," *BMC Bioinf.*, vol. 9, no. 1, 2008, Art. no. 559.
[145] M. J. Ha, V. Baladandayuthapani, and K.-A. Do, "DINGO: Differential network analysis in genomics," *Bioinf.*, vol. 31, no. 21, pp. 3413–3420, 2015.
[146] C. Siska and K. Kechris, "Differential correlation for sequencing data," *BMC Res. Notes*, vol. 10, no. 1, 2017, Art. no. 54.
[147] G. Fang, R. Kuang, G. Pandey, et al., "Subspace differential coexpression analysis: Problem definition and a general approach," in *Proc. Pacific Symp. Biocomput.*, 2010, pp. 145–56.
[148] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: A Bioconductor package for differential expression analysis of digital gene expression data," *Bioinf.*, vol. 26, no. 1, pp. 139–140, 2010.
[149] N. J. Schurch, P. Schofield, M. Gierliński, et al., "How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?" *RNA*, vol. 22, pp. 839–851, 2016.
[150] G. Finak, A. McDavid, M. Yajima, et al., "MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data," *Genome Biol.*, vol. 16, no. 1, 2015, Art. no. 278.
[151] P. S. Russo, G. R. Ferreira, L. E. Cardozo, et al., "CEMiTool: A Bioconductor package for performing comprehensive modular co-expression analyses," *BMC Bioinf.*, vol. 19, no. 1, 2018, Art. no. 56.
[152] M. Lei, J. Xu, L.-C. Huang, L. Wang, and J. Li, "Network module-based model in the differential expression analysis for RNA-seq," *Bioinf.*, vol. 33, pp. 2699–2705, 2017.
[153] O. Alter, P. O. Brown, and D. Botstein, "Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms," *Proc. Nat. Academy Sci. United States America*, vol. 100, no. 6, pp. 3351–3356, 2003.
[154] S. Ray and U. Maulik, "Identifying differentially coexpressed module during HIV disease progression: A multiobjective approach," *Sci. Rep.*, vol. 7, no. 1, 2017, Art. no. 86.
[155] H. Liany, J. C. Rajapakse, and R. K. M. Karuturi, "MultiDCoX: Multi-factor analysis of differential co-expression," *BMC Bioinf.*, vol. 18, no. 16, p. 576, 2017.
[156] J. A. Botía, J. Vandrovcova, P. Forabosco, et al., "An additional k-means clustering step improves the biological features of WGCNA gene co-expression networks," *BMC Syst. Biol.*, vol. 11, no. 1, 2017, Art. no. 47.
[157] X. Xiao, A. Moreno-Moral, M. Rotival, et al., "Multi-tissue analysis of co-expression networks by higher-order generalized singular value decomposition identifies functionally coherent transcriptional modules," *PLoS Genetics*, vol. 10, no. 1, 2014, Art. no. e1004006.
[158] J. Yang, H. Yu, B.-H. Liu, et al., "DCGL v2. 0: An R package for unveiling differential regulation from differential co-expression," *PLoS One*, vol. 8, no. 11, 2013, Art. no. e79729.
[159] B. Pontes, R. Giráldez, and J. S. Aguilar-Ruiz, "Biclustering on expression data: A review," *J. Biomed. Inform.*, vol. 57, pp. 163–180, 2015.
[160] S. de Siqueira Santos, T. F. de Almeida Galatro, R. A. Watanabe, et al., "CoGA: An R package to identify differentially co-expressed gene sets by analyzing the graph spectra," *PloS One*, vol. 10, no. 8, 2015, Art. no. e0135831.
[161] A. T. McKenzie, I. Katsyv, W.-M. Song, et al., "DGCA: A comprehensive R package for differential gene correlation analysis," *BMC Syst. Biol.*, vol. 10, no. 1, 2016, Art. no. 106.
[162] M. Padi and J. Quackenbush, "Integrating transcriptional and protein interaction networks to prioritize condition-specific master regulators," *BMC Syst. Biol.*, vol. 9, no. 1, 2015, Art. no. 80.
[163] G. Gambardella, M. N. Moretti, R. De Cegli, et al., "Differential network analysis for the identification of condition-specific pathway activity and regulation," *Bioinf.*, vol. 29, no. 14, pp. 1776–1785, 2013.
[164] D. M. Gysi, T. M. Fragoso, E. Almaas, et al., "CoDiNA: An RPackage for co-expression differential network analysis in N dimensions," arXiv:1802.00828, 2018.
[165] O. Odibat and C. K. Reddy, "Ranking differential hubs in gene co-expression networks," *J. Bioinf. Comput. Biol.*, vol. 10, no. 01, 2012, Art. no. 1240002.
[166] C.-L. Hsu, H.-F. Juan, and H.-C. Huang, "Functional analysis and characterization of differential coexpression networks," *Sci. Rep.*, vol. 5, 2015, Art. no. 13295.
[167] M. Bockmayr, F. Klauschen, B. Györffy, et al., "New network topology approaches reveal differential correlation patterns in breast cancer," *BMC Syst. Biol.*, vol. 7, no. 1, 2013, Art. no. 78.
[168] C. Gaiteri, Y. Ding, B. French, et al., "Beyond modules and hubs: The potential of gene coexpression networks for investigating molecular mechanisms of complex brain disorders," *Genes Brain Behavior*, vol. 13, no. 1, pp. 13–24, 2014.
[169] C. Gaiteri, J.-P. Guilloux, D. A. Lewis, et al., "Altered gene synchrony suggests a combined hormone-mediated dysregulated state in major depression," *PloS One*, vol. 5, no. 4, 2010, Art. no. e9970.
[170] A. Torkamani, B. Dean, N. J. Schork, et al., "Coexpression network analysis of neural tissue reveals perturbations in developmental processes in schizophrenia," *Genome Res.*, vol. 20, no. 4, pp. 403–412, 2010.
[171] T. Kakati, H. Kashyap, and D. K. Bhattacharyya, "THD-module extractor: An application for CEN module extraction and interesting gene identification for Alzheimer's disease," *Sci. Rep.*, vol. 6, 2016, Art. no. 38046.
[172] M. C. Oldham, S. Horvath, and D. H. Geschwind, "Conservation and evolution of gene coexpression networks in human and chimpanzee brains," *Proc. Nat. Academy Sci. United States America*, vol. 103, no. 47, pp. 17 973–17 978, 2006.
[173] T. W. Lui, N. B. Tsui, L. W. Chan, et al., "DECODE: An integrated differential co-expression and differential expression analysis of gene expression data," *BMC Bioinf.*, vol. 16, no. 1, 2015, Art. no. 182.

Authorized licensed use limited to: Universidad Nacional Autonoma De Mexico (UNAM). Downloaded on September 18,2024 at 20:53:55 UTC from IEEE Xplore. Restrictions apply.

[174] Y. Zuo, Y. Cui, C. Di Poto, et al., "INDEED: Integrated differential expression and differential network analysis of omic data for biomarker discovery," *Methods*, vol. 111, pp. 12–20, 2016.
[175] W. Yu, S. Zhao, Y. Wang, et al., "Identification of cancer prognosis-associated functional modules using differential co-expression networks," *Oncotarget*, vol. 8, no. 68, 2017, Art. no. 112928.
[176] S.-Q. Fang, M. Gao, S.-L. Xiong, et al., "Combining differential expression and differential coexpression analysis identifies optimal gene and gene set in cervical cancer," *J. Cancer Res. Therapeutics*, vol. 14, no. 1, pp. 201–207, 2018.
[177] M. Drag, R. Skinkyté-Juskiené, D. N. Do, et al., "Differential expression and co-expression gene networks reveal candidate biomarkers of boar taint in non-castrated pigs," *Sci. Rep.*, vol. 7, no. 1, 2017, Art. no. 12205.
[178] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists," *Nucleic Acids Res.*, vol. 37, no. 1, pp. 1–13, 2008.
[179] G. F. Berriz, J. V. White, O. D. King, et al., "GoFish finds genes with combinations of Gene Ontology attributes," *Bioinf.*, vol. 19, no. 6, pp. 788–789, 2003.
[180] B. R. Zeeberg, W. Feng, G. Wang, et al., "GoMiner: A resource for biological interpretation of genomic and proteomic data," *Genome Biol.*, vol. 4, no. 4, 2003, Art. no. R28.
[181] F. Al-Shahrour, R. Díaz-Uriarte, and J. Dopazo, "FatiGO: A web tool for finding significant associations of Gene Ontology terms with groups of genes," *Bioinf.*, vol. 20, no. 4, pp. 578–580, 2004.
[182] B. Zhang, D. Schmoyer, S. Kirov, and J. Snoddy, "GOTree machine (GOTM): A web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies," *BMC Bioinf.*, vol. 5, no. 1, 2004, Art. no. 16.
[183] S. Volinia, R. Evangelisti, F. Francioso, et al., "GOAL: Automated Gene Ontology analysis of expression profiles," *Nucleic Acids Res.*, vol. 32, no. suppl_2, pp. W492–W499, 2004.
[184] G. F. Berriz, O. D. King, B. Bryant, et al., "Characterizing gene sets with FuncAssociate," *Bioinf.*, vol. 19, no. 18, pp. 2502–2504, 2003.
[185] S. W. Doniger, N. Salomonis, K. D. Dahlquist, et al., "MAPPFinder: Using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data," *Genome Biol.*, vol. 4, no. 1, 2003, Art. no. R7.
[186] M. Masseroli, D. Martucci, and F. Pinciroli, "GFINDer: Genome function integrated discoverer through dynamic annotation, statistical analysis, and mining," *Nucleic Acids Res.*, vol. 32, no. suppl_2, pp. W293–W300, 2004.
[187] X. Yang, J. Li, Y. Lee, et al., "GO-Module: Functional synthesis and improved interpretation of Gene Ontology patterns," *Bioinf.*, vol. 27, no. 10, pp. 1444–1446, 2011.
[188] E. Glaab, A. Baudot, N. Krasnogor, et al., "EnrichNet: Network-based gene set enrichment analysis," *Bioinf.*, vol. 28, no. 18, pp. i451–i457, 2012.
[189] M. Kanehisa and S. Goto, "KEGG: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, 2000.
[190] D. Nishimura, "BioCarta," *Biotech Softw. Internet Rep.: Comput. Softw. J. Sci.*, vol. 2, no. 3, pp. 117–120, 2001.
[191] P. Creixell, J. Reimand, S. Haider, et al., "Pathway and network analysis of cancer genomes," *Nature Methods*, vol. 12, no. 7, 2015, Art. no. 615.
[192] L. Zhang and S. Zhang, "Comparison of computational methods for imputing single-cell RNA-sequencing data," *IEEE/ACM Trans. Comput. Biology Bioinf.*, 2018.
[193] P. Mahanta, H. A. Ahmed, D. K. Bhattacharyya, and J. K. Kalita, "An effective method for network module extraction from microarray data," *BMC Bioinf.*, vol. 13, no. 13, 2012, Art. no. S4.
[194] H. A. Ahmed, P. Mahanta, D. K. Bhattacharyya, and J. K. Kalita, "Shifting-and-scaling correlation based biclustering algorithm," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 11, no. 6, pp. 1239–1252, Nov./Dec. 2014.
[195] P. Sharma, D. K. Bhattacharyya, and J. Kalita, "Disease biomarker identification from gene network modules for metastasized breast cancer," *Sci. Rep.*, vol. 7, no. 1, 2017, Art. no. 1072.
[196] K. Mandal, R. Sarmah, and D. K. Bhattacharyya, "Biomarker identification for cancer disease using biclustering approach: An empirical study," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 2, pp. 490–509, Mar./Apr. 2019.
[197] R. Uddin and S. M. Singh, "Gene network construction from microarray data identifies a key network module and several candidate hub genes in age-associated spatial learning impairment," *Frontiers Syst. Neurosci.*, vol. 11, 2017, Art. no. 75.

Hussain Ahmed Chowdhury received the BE degree in computer science & engineering from Jorhat Engineering College, in 2013, and the MTech degree in information technology from Tezpur University, in 2016. Currently, he is working toward the PhD degree in computer science at Tezpur University. His current research interests include machine learning and computational biology.

Dhruba Kumar Bhattacharyya received the PhD degree in computer science from Tezpur University, in 1999. He is a professor with the Computer Science & Engineering Department, Tezpur University. His research areas include data mining, network security, and content-based image retrieval. He has published more than 250 research papers in the leading international journals and conference proceedings. In addition, he has written/edited many books. He is a program committee/advisory body member of several international conferences/workshops.

Jugal Kumar Kalita received the PhD degree from the University of Pennsylvania, in 1990. He is a professor of computer science, University of Colorado at Colorado Springs. His research interests include natural language processing, artificial intelligence, and bioinformatics. He has published more than 200 papers in international journals and referred conferences.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/csdl.

Authorized licensed use limited to: Universidad Nacional Autonoma De Mexico (UNAM). Downloaded on September 18,2024 at 20:53:55 UTC from IEEE Xplore. Restrictions apply.