Cellular and Molecular Biology
E-ISSN : 1165-158X / P-ISSN : 0145-5680
www.cellmolbiol.org

Original Research

ARNetMiT R Package: association rules based gene co-expression networks of miRNA targets

M. Özgür Cingiz*, G. Biricik, B. Diri

Computer Engineering Department, Yildiz Technical University, Istanbul, Turkey

Correspondence to: mozgur@ce.yildiz.edu.tr

Received November 29, 2016; **Accepted** March 25, 2017; **Published** March 31, 2017

Doi: http://dx.doi.org/10.14715/cmb/2017.63.3.4

Abstract: miRNAs are key regulators that bind to target genes to suppress their gene expression level. The relations between miRNA-target genes enable users to derive co-expressed genes that may be involved in similar biological processes and functions in cells. We hypothesize that target genes of miRNAs are co-expressed, when they are regulated by multiple miRNAs. With the usage of these co-expressed genes, we can theoretically construct co-expression networks (GCNs) related to 152 diseases. In this study, we introduce ARNetMiT that utilize a hash based association rule algorithm in a novel way to infer the GCNs on miRNA-target genes data. We also present R package of ARNetMiT, which infers and visualizes GCNs of diseases that are selected by users. Our approach assumes miRNAs as transactions and target genes as their items. Support and confidence values are used to prune association rules on miRNA-target genes data to construct support based GCNs (sGCNs) along with support and confidence based GCNs (scGCNs). We use overlap analysis and the topological features for the performance analysis of GCNs. We also infer GCNs with popular GNI algorithms for comparison with the GCNs of ARNetMiT. Overlap analysis results show that ARNetMiT outperforms the compared GNI algorithms. We see that using high confidence values in scGCNs increase the ratio of the overlapped gene-gene interactions between the compared methods. According to the evaluation of the topological features of ARNetMiT based GCNs, the degrees of nodes have power-law distribution. The hub genes discovered by ARNetMiT based GCNs are consistent with the literature.

Key words: Gene co-expression network; Association rule based algorithms; GNI algorithms; miRNA-target genes.

Introduction

Advances in sequencing technologies equip researchers for further investigation and understanding of underlying mechanisms in disorders. Sequencing techniques like microarray gene expression data, RNA-sequencing data (RNA-seq), ChIP-sequencing (ChIP-seq), microRNA (miRNA) give detailed overview of entire genomes and transcriptomes. These high-throughput techniques reveal molecular interactions and allow them to be represented as networks. In these networks, molecules and their interactions are represented as nodes and edges. We know that molecular relations can take active role in regulation of biological processes that are related to the pathogenesis of cancer. For this reason, understanding the structure of molecular interactions is the key to reveal the cause of disorders.

miRNAs are small non-coding RNA molecules that bind to messenger RNA (mRNA) transcripts and post-transcriptionally regulate expression of target genes. Cell growth, differentiation, proliferation, apoptosis, migration and similar processes are associated with cancer. miRNAs play important role in regulation of these processes by controlling expression of target genes. The inferred networks of miRNAs and target genes are involved in many cancer related biological processes. Recent studies (1-3) use miRNA expression data to infer co-expression networks and regulatory networks. Information based Network Inference (NI) (1), Bayesian Networks (2), Differential Equation (3) are some of the popular methods that use miRNA expression data for inferring gene networks.

In this study, CoMeTa (4) project inspired us to derive gene networks from miRNA-target genes dataset. CoMeTa infers GCNs by using miRNA-target genes data and microarray gene expression data. First, target genes of miRNAs are retrieved. Afterwards, the relation between the target genes and other genes are obtained from microarray gene expression data to construct the gene co-expression networks.

Co-expressed genes are involved in similar biological processes and functions in the cells. Identification of co-expressed genes is important since Transcription Factors (TFs) and miRNAs have tendency to regulate the co-expressed genes. This point motivates us to infer gene co-expression networks of miRNA targets that are related to different disorders. In this study, we use the experimentally validated interactions of miRNA targets for GCN construction. We hypothesize that target genes of miRNAs are co-expressed, when they are regulated by multiple miRNAs. In addition, the strength of the correlation between co-expressed genes increase when the number of common regulating miRNAs increase.

It is known that association rule mining algorithms can detect relations in large datasets and can specify most relevant items (5-8). In addition, they serve for constructing graphs with these relevant items. For this reason, we use a hash based association rule algorithm to infer gene-gene interactions from miRNA-target genes. In the initial phase, we assume miRNAs as tran-

sactions, and the targets of miRNAs as items. Next, we use support and confidence values to extract association rules from miRNA-target genes data. Finally, we estimate co-expressed genes on the validated miRNA-target genes relations. We evaluate the proposed approach through overlap analysis of the gene network with literature data, which consist of Protein-Protein Interactions (PPI). Additionally, we examine topological features of the inferred gene networks and the relations between hub nodes of networks and cancer types. Using the miRNA-target gene relations, we compared the similarities between the GCNs of our proposed approach with the GCNs of information based gene network inference (GNI) algorithms. Besides these similarities that give information about the shared relations between the derived GCNs, we also compared their performances on validation data.

In the next Section, we introduce the dataset and the association rule mining method. Section 3 presents the experimental results. Finally, Section 4 discusses the results and concludes the paper.

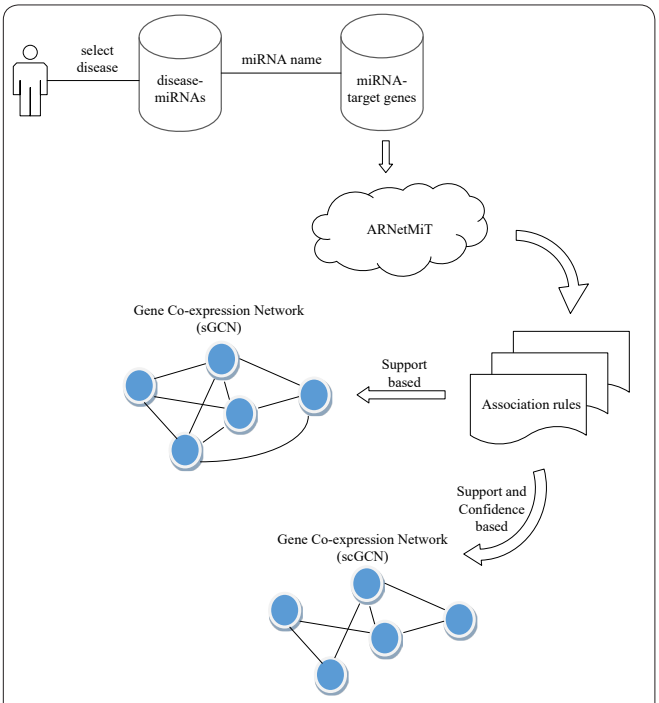

Figure 1. The overview of the proposed ARNetMiT system.

Materials and Methods

We use a hash based association rule mining algorithm in a novel way to infer the GCNs on miRNA-target genes data. In this section, we introduce our dataset and our method.

Dataset

We obtain miRNA relations for diseases from the manually curated miR2Disease database (9). The miR2Disease database contains 2,429 unique miRNA-disease relations, associated with 477 miRNAs and 152 diseases.

We gather the miRNA-target gene relations from two sources. The first resource is the 6.1 release of the miRTarBase (10) database that contains experimentally validated 322,389 unique miRNA-target gene interactions, between 2,649 miRNAs and 14,894 target genes. Second resource is again the miR2Disease database, containing 637 unique miRNA-target gene instructions between 180 miRNAs and 397 target genes.

Our integrated dataset consists of experimentally validated 322,994 unique miRNA-target gene interactions. These interactions are related to 2,801 miRNAs and 15,059 target genes. We generate the disease related miRNA-target genes data from the relations between disease-miRNAs and miRNA-target genes. We match these relations based on miRNA names. After data retrieval, we apply our association rule based algorithm to construct GCNs.

Association rules based gene network of miRNA targets (ARNetMiT)

We aim to construct disease related co-expression gene network on miRNA-target genes data. We also give users the option of selecting any disease to get the related GCN. We name the proposed model as Association Rules based Gene Network of miRNA Targets (ARNetMiT). The overview on the design of ARNetMiT is given in Figure 1.

Association rule based algorithms work effectively on discovering hidden relations between items, especially in big datasets. Apriori (11) and Eclat (12) are the most popular and widely used association rule based algorithms and they are applied in many different areas. These two algorithms differ on search methods when traversing a tree. While Apriori traverses a prefix tree with breath first search, Eclat favors depth first search. Because of this difference, Apriori has a disadvantage of scanning the dataset multiple times. Eclat prevents the multiple scanning problem by intersecting rows to discover hidden relations in the itemset (13). In this study, we modified the hash based Eclat algorithm and used in GCN construction. Since our dataset consists of miRNA-target gene relations, Eclat transposes the data to find the common miRNAs that regulate the co-expressed genes. We solve the excessive size problem of the integrated miRNA-target genes data by using a hash function. The hash table produced by the hash function enables us to rapidly retrieve data of the candidate items.

In order to eliminate insignificant interactions, association rule based algorithms use several parameters such as support, confidence and lift. In this study we use support value, which is calculated by (1) and (2), to detect the frequent genes that are regulated by the same miRNAs. We find the frequent genes that are regulated by more than a number of miRNAs with (1). The minimum number of regulating miRNAs are determined using a threshold value ($minsup_1$). In (1), G is the set of all genes, $n(x)$ is the number of miRNAs that regulate the gene x, N is the total number of miRNAs and F_1 is the frequent 1-itemset of genes.

$$\text{support}(x) = \frac{n(x)}{N}$$
$$\forall x \in G : \{\text{support}(x) > minsup_1 \rightarrow x \in F_1\} \quad (1)$$

Candidates of co-expressed genes are the pairs of frequent genes in F_1. We find the co-expressed genes that are regulated by more than a number of common miRNAs with (2). In this equation, $\{x,y\}$ gene pair is a candidate of co-expressed genes and F_2 is the frequent

2-itemset of gene pairs. The minimum number of miRNAs ($minsup_2$) that regulate $\{x,y\}$ are determined using the same thresholding in (1).

$$\forall\{x,y\}\in F_1 : \left\{\frac{n(x,y)}{N} > minsup_2 \rightarrow \{x,y\}\in F_2\right\} \quad (2)$$

Higher support value for the co-expressed genes implies that there are many common miRNAs regulating both genes. Thus, higher support value drives us to say that both genes have tendency to be co-expressed. The minimum support value (*minsup*) is a mandatory parameter of ARNetMiT, including the options of first quartile, median, mean and third quartile values of the patterns' abundance distributions. The default *minsup* choice is rank support type, where the user defined coefficient is multiplied by the maximum abundance value of the patterns. The $minsup_1$ and $minsup_2$ values are calculated in the same way for F_1 and F_2 respectively.

Association rule mining algorithms can offer various parameters in addition to support value to prune weak hidden relations. Confidence (3) is such a parameter to eliminate weak gene-gene relations. The confidence value defines the conditional dependencies between co-expressed genes that are regulated by common miRNAs. According to (3), when gene x is given, the conditional probability of gene y is calculated with the division of the number of miRNAs that regulate the gene pair $\{x,y\}$ to the number of miRNAs that regulate x. If the conditional probability approximates to 1, we can infer that x and y are likely co-expressed genes, when x is given. The *minconf* value in (3) is a user defined parameter and F_c is the set of gene-gene interactions whose support and confidence values are higher than *minsup* and *minconf*.

$$\text{confidence}(y \Rightarrow x) = P(y \mid x) = \frac{n(x,y)}{n(x)} \quad (3)$$
$$\forall x,y \in F_1 \wedge \forall\{x,y\}\in F_2 : \{\text{confidence}(y \Rightarrow x) > minconf \rightarrow \{x,y\}\in F_c\}$$

Example: Consider the following association rules, derived from F_1 = {$Gene_1$, $Gene_2$, $Gene_3$, $Gene_4$, $Gene_5$} and $minsup_2$=0.5, *minconf*=0.5.
$rule_1$: (P($Gene_1$|$Gene_2$) > $minsup_2$), then $Gene_2 \Leftrightarrow Gene_1$
$rule_2$: (P($Gene_2$|$Gene_3$) > $minsup_2$), then $Gene_3 \Leftrightarrow Gene_2$
$rule_3$: (P($Gene_4$|$Gene_1$) > $minsup_2$), then $Gene_1 \Leftrightarrow Gene_4$
$rule_4$: (P($Gene_5$|$Gene_1$) > $minsup_2$), then $Gene_1 \Leftrightarrow Gene_5$

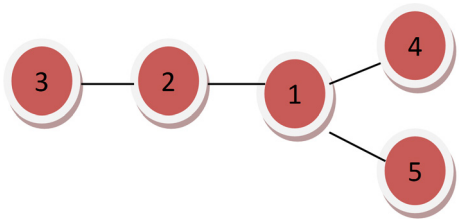

Figure 2. Example of sGCN obtained by using the support value.

$rule_1$: (P($Gene_1$|$Gene_2$) > $minsup_2$ & P($Gene_1$|$Gene_2$) > *minconf*), then $Gene_2 \Leftrightarrow Gene_1$
$rule_2$: (P($Gene_2$|$Gene_3$) > $minsup_2$ & P($Gene_2$|$Gene_3$) > *minconf*), then $Gene_3 \Leftrightarrow Gene_2$
$rule_3$: (P($Gene_4$|$Gene_1$) > $minsup_2$ & P($Gene_4$|$Gene_1$) > *minconf*), then $Gene_1 \Leftrightarrow Gene_4$
$rule_4$: (P($Gene_5$|$Gene_1$) > $minsup_2$ & P($Gene_5$|$Gene_1$) ≤ *minconf*), then {$Gene_1$, $Gene_5$} is not co-expressed.

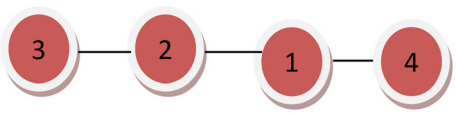

Figure 3. Example of scGCN obtained by using the support and the confidence values.

Figure 2 shows an example sGCN illustration, obtained by using only the support value. The scGCN in Figure 3 is obtained by using the support and the confidence values. ARNetMiT builds graphs as in the examples given in Figures 2 and 3. By using these graphs, ARNetMiT can theoretically derive GCNs of 152 diseases.

The steps of ARNetMiT are comprehensively illustrated in Figure 4. The first part of the illustration describes the miRNA-target gene relations. We transformed the raw binary miRNA-target genes data into matrix format in order to proceed to the calculation of the most regulated genes.

In the second part, we introduce the gene occurrences, which describe the number of miRNAs regulating each gene. Since the total number of miRNAs are same for each gene, the denominators (N) of (1) and (2) can be omitted. The minimum support value for the example in Figure 4 is 2 ($minsup_1$=0.5, maximum occurrence value (*MOV*)=4). In this example, we removed the genes having less than 3 regulators from F_1.

The third part of the illustration shows the co-expressed genes, which consist of possible interactions of the genes in F_1. For instance, this part shows that $Gene_B$ and $Gene_C$ are regulated by 3 common miRNAs. At the next step, we use the minimum support value ($minsup_2$) to prune the gene-gene interactions. Similar to 1-itemset pruning, the $minsup_2$ value for F_2 in Figure 4 is 1.5 ($minsup_2$=0.5, *MOV*=3). At this step ARNetMiT produces sGCN.

In our example, if a gene-gene interaction is regulated by at least 2 common miRNAs, we assume that them as co-expressed genes. We calculate confidence values of these co-expressed genes by using (3). We eliminate the co-expressed genes having confidence value below *minconf* (in this example, 0.5). At the last step, we make the gene-gene interaction matrix symmetric to form the scGCNs of ARNetMiT.

At this point, we introduce ARNetMiT R package, which enables the users to infer sGCNs and scGCNs. The datasets mentioned in Section 2.1 are also given in the ARNetMiT R package. ARNetMiT can give the lift values of gene-gene pairs to calculate the relevance score of gene-gene interactions. The package can also list the all regulator miRNAs of the co-expressed genes. ARNetMiT uses RedeR R package (14) to visualize the GCNs. Researchers can access ARNetMiT R Package and the user manual from https://sites.google.com/site/arnetmit/.

Results

For the performance analysis of ARNetMiT GCNs, we use overlap analysis, GNI network comparison, and

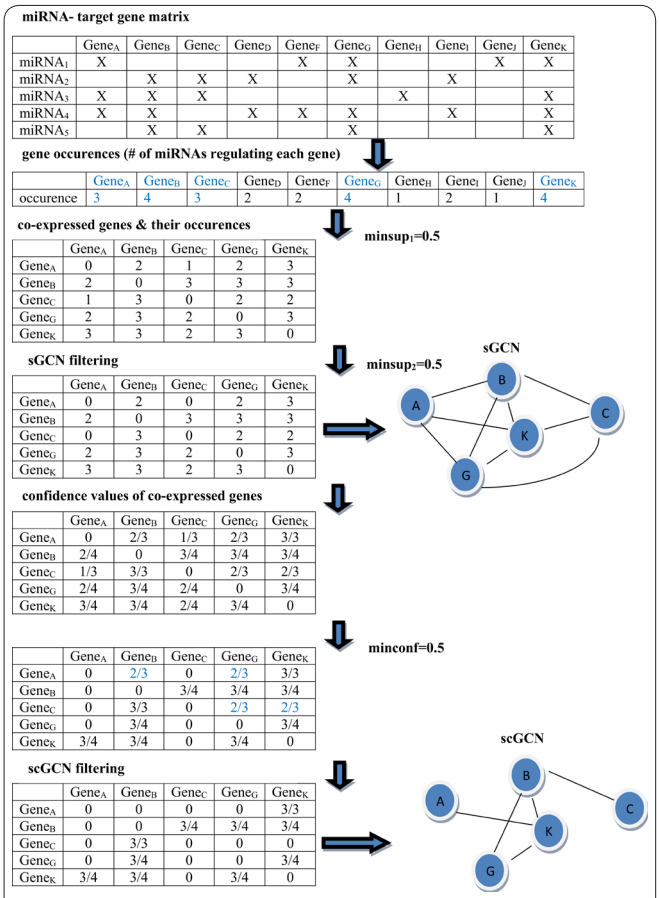

Figure 4. The main steps of ARNetMiT, explained with an example.

topological features of gene networks. First, we focus on overlap analysis using the literature data. Then we compare the GCNs of ARNetMiT with the GCNs of GNI algorithms. Later on, we examine the topological features of ARNetMiT.

Overlap analysis of ARNetMiT based gene networks using the literature data

In overlap analysis, previous studies use synthetic and real biological datasets for performance analysis of the gene networks (15, 16). PPI datasets that consist of interactions are frequently used in overlap analysis with real biological data (17, 18). Thus, we use PPI datasets for evaluating the performance of GCNs. The validation data consists of 1,594,366 unique biologically validated interactions. This number creates high false negative rates in the results of gene network inference (GNI) algorithms. For this reason, we use both number of true positives (TP) and precision in performance analysis. Afterwards, we utilize Fisher's Exact Test (FET) to determine whether our overlap analysis is statistically significant or not. We discard all gene interactions that have significance lower than 0.05. We use GAnet R package (19) for the precision and p-value calculations.

We utilize data associated with breast, colorectal, pancreatic and prostate cancers, which are derived from miRNA-target genes data of ARNetMiT R package in order to construct the GCNs. We choose $minsup_{1,2}$={0.1, 0.2, 0.3} for sGCN inference using the rank type parameter. The result patterns for each of the four cancers are similar. When the support value is 0.3, precisions of all selected cancer related GCNs are higher, which can be observed in Figure 5. Although the precision is high, the true positive rates are lower as the number of interactions decrease. However, these strong association rules lead us to derive more accurate and robust gene-gene interactions. Figure 5 shows that ARNetMiT scores nearly 10,000 TPs when *minsup*=0.1 but the precision is below 0.1. This result implies that ARNetMiT predicts many false positives when the support value is chosen low. The inverse proportion between precision and number of TPs is valid for all GCNs.

As we mentioned before, ARNetMiT enables users to utilize both support and confidence values to eliminate the insignificant association rules, when constructing the GCNs. After the support value, the confidence value provides a second pruning step for weak associations. Since higher support values provide strong filtering, we choose *minsup*=0.1 and *minconf*={0.5, 0.7, 0.9} in order to construct scGCNs.

The same literature data, which we used in performance evaluation of sGCNs, is used for overlap analysis of scGCNs. Figure 6 proves that lower minimum confidence values do not produce lower precisions. This is different from the relation between support values and precisions in sGCNs. Precisions of scGCNs obtained with lower *minconf* values are generally higher than or equal to the precisions acquired with higher *minconf* values. On the other hand, number of TPs and interaction predictions between sGCNs and scGCNs are similar when higher thresholds are chosen. Thus, lower *minconf* values produce higher TPs, as a result of low pruning. When we observe Figure 6, we see that the precisions and TPs of sGCNs and scGCNs are very close. However, the ability of using both support and confidence values in ARNetMiT enables users to fine tune the parameters for any data.

We compare the performance of GCNs obtained with ARNetMiT and popular information based GNI

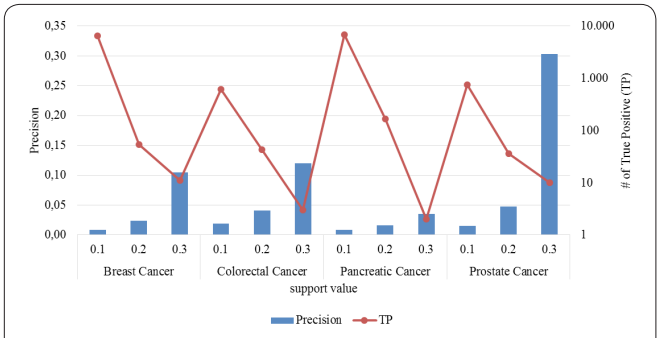

Figure 5. Precisions and the number of TPs of four cancer related sGCNs.

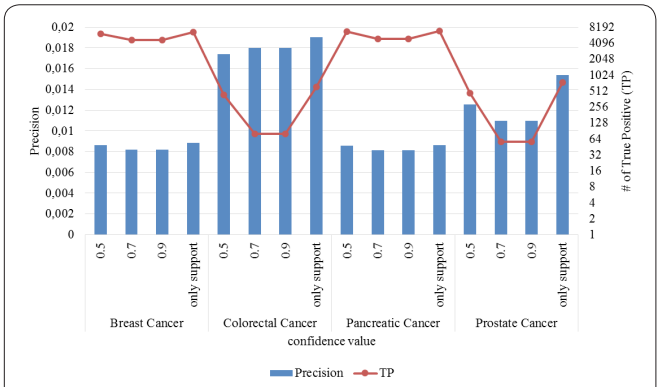

Figure 6. Precisions and the number of TPs of four cancer related scGCNs.

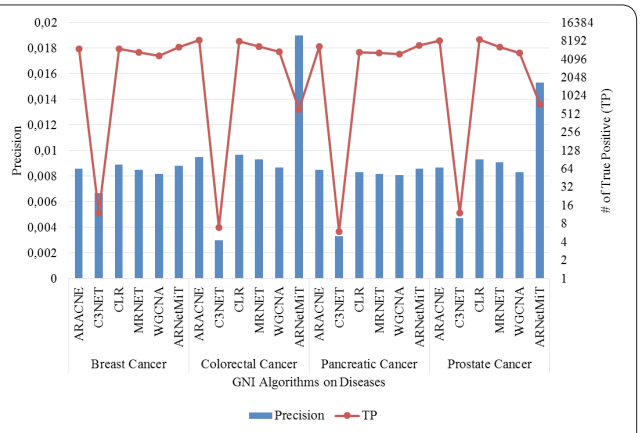

Figure 7. Performance comparison of ARNetMiT and GNI algorithms on four types of cancer.

algorithms using the same miRNA-target genes data. The *buildmiRNATargetTable* function of ARNetMiT enables us to transform miRNA-target genes data to the format that the GNI algorithms use. Among the popular GNI algorithms, we choose ARACNE (20), C3NET (21), CLR (22), MRNET (23), and WGCNA (24) for the comparison. The minet R package (25) provides ARACNE, CLR and MRNET algorithms. Figure 7 presents the comparison of the GCNs obtained with the selected GNI algorithms and the GCN of ARNetMiT (a sGCN with *minsup*=0.1) on four cancer types. CLR outperforms other GNI algorithms on breast, colorectal and prostate cancers. The highest precision of GNI algorithms is obtained by ARACNE on prostate cancer. On the other hand, GCNs of ARNetMiT produce slightly higher precision values than GNI algorithms on breast and pancreatic cancers. Additionally, ARNetMiT significantly outperforms the compared methods on colorectal and prostate cancers. Besides the superiority on precision, the number of gene-gene interactions inferred by ARNetMiT is very close to the compared methods, which makes it preferable for the researchers.

We investigate the overlapped interactions to compare the similarities between gene-gene interactions of the GCNs obtained with ARNetMiT and GNI algorithms. On this analysis, we used three GCNs of ARNetMiT, which are inferred with {*minsup*=0.1, *minsup*=0.1 & *minconf*=0.5, *minsup*=0.1 & *minconf*=0.7} and compared them with the most successful GNI method on each of the four cancer types. Figure 8 shows that the estimated interactions of ARNetMiT and the compared algorithms are closely overlapped, when both *minsup* and *minconf* filters are employed. A closer inspection of Figure 8 reveals the sensitivity of choosing filter values in ARNetMiT. When we use only *minsup*, the overlap ratio is at minimum. Integrating a lower *minconf* value enhances the ratio. However, the usage of *minsup* with higher values of *minconf* boosts up the overlapping gene-gene interactions.

Topological features of ARNetMiT based gene networks

Biological networks are scale-free networks, where the nodes approximate power-law degree distribution. These networks contain a few number of hub nodes that have significantly more connections than other nodes. In gene networks, hub genes are involved in many biological processes associated with cancers. For this reason, it is crucial to identify these hub genes in order to discover the disease related genes. Besides hub genes, the structural features of gene networks are supplementary elements to measure the fitness of GCNs to scale-free networks. For this topological assessment, we use three parameters. The first one is the average number of neighbors for the nodes. This parameter indicates interconnectivity of genes. Network heterogeneity (NH) is the second parameter and measures the variance of node degrees. High heterogeneity implies whether the network is relevant to the power-law degree distribution or not. The third parameter is clustering coefficient (CC), which gives information about the clustering tendency of nodes. We use NetworkAnalyzer tool in Cytoscape (26) to obtain these three topological parameters. The GCNs are evaluated with respect to the topological features with *minsup*=0.2. This support value supplies moderate balance between number of predictions and the precision.

Table 1 shows the topological features of ARNetMiT based GCNs. In the table, the first four rows report the sGCNs that are inferred with *minsup*=0.2. The last four rows of Table 1 list the scGCNs, inferred using *minsup*=0.2 and *minconf*=0.5.

In biological networks with power-law degree distribution, NH and CC values are close to 1. These values are close to 0 in random biological networks. Table 1 indicates that the nodes of networks are tightly coupled, having high average number of neighbor nodes. Maximum of this parameter is for the sGCN of pancreatic cancer, also approved by the highest clustering coefficient value of 0.945. In biological networks, high average number of neighbors are in direct proportion to high clustering coefficient values. The nodes of all GCNs have high CC values that emphasize their clustering tendency. Although two-step pruning in scGCNs decreases the average number of neighbors of nodes, it increases NH and CC values, due to the elimination of weak associations between genes.

The hub genes of GCNs derived only by degree of nodes are listed in the last column of Table 1. Hub genes of sGCNs and scGCNs of breast cancer are same except one gene. Studies prove that XIAP and PARD6B genes are involved in breast cancer related processes and their

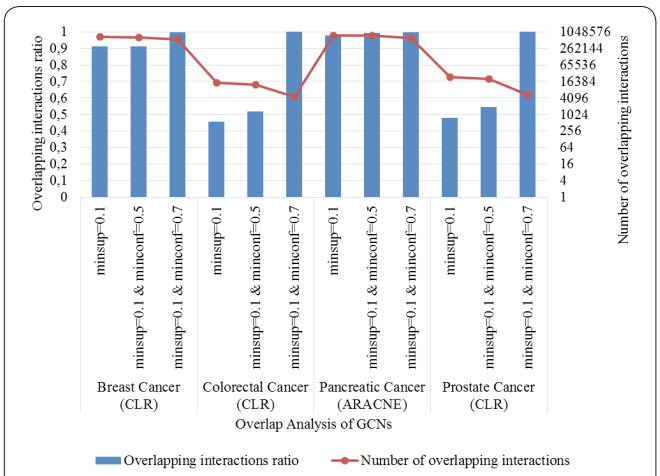

Figure 8. Comparison of overlapping interactions between ARNetMiT and best performing GNI algorithm on four types of cancer.

Table 1. Topological features and hub genes of ARNetMiT based GCNs.

Gene Network & types	Average # of Neighbors	Network Heterogeneity	Clustering Coefficient	Hub Genes
Breast Cancer, sGCN	25.141	0.802	0.868	NUFIP2, XIAP, PARD6B, PRRG4, MYC, BLC2
Colorectal Cancer, sGCN	34.033	0.417	0.809	GLO1, BCL2, GIGYF1, NCOA3, PAIP1, ITGA2
Pancreatic Cancer, sGCN	124.5	0.346	0.945	BCL2, PLAG1, STX6, RAB10, PTEN, GATA6
Prostate Cancer, sGCN	20.8	0.665	0.73	BCL2, SF3B3, GIGYF1, CPOX, TP53, GRPEL2
Breast Cancer, scGCN	24.786	0.814	0.872	NUFIP2, XIAP, PARD6B, PRRG4, MYC, MLLT1
Colorectal Cancer, scGCN	12.417	0.721	0.721	NR2F6, THBS1, NUFIP2, NAP1L1, SF3B3, WDR82
Pancreatic Cancer scGCN	114.122	0.398	0.924	TP53, MYC, ELK4, PER1, ASH1L, HIF1A
Prostate Cancer, scGCN	19.043	0.721	0.732	SF3B3, GIGYF1, ACSL4, YWHAZ, RAB10, NRBP1

mutations cause breast cancer (27, 28). BCL2 and MYC are apoptosis associated and regulator genes in many of the tumor cells that belong to a variety of cancers (29, 30). The role of ITGA2 polymorphism and the overexpression of GLO1 gene in tumor cells are important issues in colorectal cancer (31, 32). The overexpression of NR2F6 and NAP1L1 are associated to colorectal cancer (33, 34). The overexpression of PLAG1, STX6 genes and low-expression of ELK4 gene in tumor cells are associated to pancreatic cancer (35, 36). YWHAZ and ACSL4 genes promote prostate cancer and they are defined as prostate cancer biomarkers (37, 38). For all of these four cancers, the hub genes of gene networks inferred by ARNetMiT are related to the genes mentioned above. This shows that the hub genes discovered by ARNetMiT based gene networks are consistent with the literature.

Discussion

The main objective of ARNetMiT is to construct disease related GCNs. In order to achieve this, a hash based association rule algorithm is used to find the hidden relations that reveal the co-expressed genes. When the user chooses the disease type in ARNetMiT, miRNA-target genes are determined using miRNA names, which are the pairing keys between disease-miRNA and miRNA-target genes data. The disease-miRNAs and miRNA-target genes data are both validated. Hence, the results do not contain any noisy samples. We assume miRNAs as transactions and genes as items. ARNetMiT uses support and confidence values to extract the association rules, which show the gene-gene interactions. Users have two options for building GCNs. ARNetMiT produces sGCNs when only support value is used. Utilizing support and confidence values together results in scGCNs. The R package for ARNetMiT enables users to construct GCNs of selected diseases.

Previous studies use the miRNA expression data to infer GCNs by using graphical models and information based GNI algorithms. Our contribution is to construct GCNs of 152 different diseases on validated miRNA-target genes data by using hash based association rule algorithm. The R package of ARNetMiT provides the visualization of the GCNs to the users. Besides its advantages, the drawback of ARNetMiT is parameter selection, as it works relatively slow with lower support and confidence values. This is due to the computational complexity of the association based algorithm.

The results of ARNetMiT emphasize the efficiency of our approach. Initial evaluation is carried through the overlap analysis with literature data. We labeled the overlapping interactions of ARNetMiT and literature data as true positives. When the literature data contains over 1 million gene-gene interactions, false negatives in gene networks are high. For this reason, we use precision and p-values of FET in performance evaluation.

Precisions of sGCNs and scGCNs vary between 0.05 and 0.30 on PPI data, which resemble the precisions reported before (17, 18). In sGCNs, lower *minsup* values increase the number of gene-gene interactions. However, these associations are weak and produce lower precision. When less interactions are predicted by using high *minsup* values, precisions of sGCNs increase. However, this inverse proportion is not exactly the same for scGCNs. Two-step pruning by using both support and confidence values in scGCNs have minor impact on precision.

We compared the GCNs of ARNetMiT with the GCNs inferred by popular GNI algorithms on the same miRNA-target genes data of four cancer types. ARNetMiT is able to transform miRNA-target genes data to the format that the GNI algorithms use. This function enables us to compare the methods, which is performed with overlap analysis on literature data. Our results prove that GCNs of ARNetMiT outperforms GNI based GCNs. The overlap analysis of these GCNs reveals that ARNetMiT and GNI algorithms predict many common gene-gene interactions. The interactions predicted with ARNetMiT and GNI algorithms become identical when pruning in ARNetMiT increase. This situation is clearly visible in scGCNs that use two step pruning with high confidence values.

Our second performance evaluation of ARNetMiT based GCNs is based on their topological features. As we mentioned before, biological networks are scale free

and the degrees of nodes show power-law distribution. In power-law distribution, degrees of the hub genes are significantly higher than the degrees of the remaining nodes. Hub genes, which are few in number, regulate many cancer related biological processes. We see that the hub genes of ARNetMiT based GCNs are consistent with the previous studies. The overexpression or mutations of BCL2, MYC, TP53, and RAB10 genes are introduced as cancer biomarkers in the literature (29, 30, 39) and ARNetMiT also found these hub genes.

In contrast to random networks, biological networks have higher clustering coefficient and network heterogeneity values that approximate to 1. The clustering coefficient scores of ARNetMiT based GCNs emphasize that the nodes of inferred GCNs are modular and highly coupled. The degree of these nodes show power-law distribution, since their network heterogeneity values approximate to 1. Thus, our topological features fit the topological features of biological networks.

We used literature data and topological features of gene networks to evaluate their performances. We plan to use RNA-seq data and microarray gene expression data for the performance comparison of ARNetMiT based GCNs in our future work.

References

1. Mamdani M, Williamson V, McMichael GO, Blevins T, Aliev F, Adkins A, et al. Integrating mRNA and miRNA Weighted Gene Co-Expression Networks with eQTLs in the Nucleus Accumbens of Subjects with Alcohol Dependence. PloS ONE 2015; 10(9): e0137671.
2. Liu B, Li J, Tsykin A, Liu L, Gaur AB, Goodall GJ. Exploring complex miRNA-mRNA interactions with Bayesian networks by splitting-averaging strategy. BMC Bioinformatics 2009; 10(1): 408.
3. Lai X, Bhattacharya A, Schmitz U, Kunz M, Vera J, Wolkenhauer O. A systems' biology approach to study microRNA-mediated gene regulatory networks. BioMed Res Int 2013; 2013: 703849.
4. Gennarino VA, D'Angelo G, Dharmalingam G, Fernandez S, Russolillo G, Sanges R, et al. Identification of microRNA-regulated gene networks by expression analysis of target genes. Genome Res 2012; 22(6): 1163-72.
5. Chawla S, Joseph GD, Pandey G. On Local Pruning of Association Rules Using Directed Hypergraphs. Proc ICDE'04 2004; 832.
6. Yang DH, Kang JH, Park YB, Park YJ, Oh HS, Kim SB. Association rule mining and network analysis in oriental medicine. PLoS ONE 2013; 8(3): e59241.
7. Martínez-Ballesteros M, Nepomuceno-Chamorro IA, Riquelme JC. Discovering gene association networks by multi-objective evolutionary quantitative association rules. J Comput Syst Sci 2014; 80(1): 118-36.
8. Belyi E, Giabbanelli PJ, Patel I, Balabhadrapathruni NH, Abdallah AB, Hameed W, et al. Combining association rule mining and network analysis for pharmacosurveillance. J Supercomput 2016; 72(5): 2014-34.
9. Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, et al. miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic Acids Res 2009; 37(Suppl.1): D98-104.
10. Chou CH, Chang NW, Shrestha S, Hsu SD, Lin YL, Lee WH, et al. miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. Nucleic Acids Res 2016; 44(D1): D239-47.
11. Agrawal R, Ramakrishnan S. Fast algorithms for mining association rules. Proc VLDB'94 1994; 487-99.
12. Zaki MJ. Scalable algorithms for association mining. IEEE T Knowl Data En 2000; 12(3): 372-90.
13. Borgelt C. Efficient implementations of apriori and eclat. Proc FIMI'03 2003; 26-34.
14. Castro MA, Wang X, Fletcher MN, Meyer KB, Markowetz F. RedeR: R/Bioconductor package for representing modular structures, nested networks and multiple levels of hierarchical associations. Genome Biol 2012; 13(4): R29.
15. Van den Bulcke T, Van Leemput K, Naudts B, van Remortel P, Ma H, Verschoren A, et al. SynTReN: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. BMC Bioinformatics 2006; 7(1): 43.
16. Altay G. Empirically determining the sample size for large-scale gene network inference algorithms. IET Syst Biol 2012; 6: 35-43.
17. Giorgi FM, Del Fabbro C, Licausi F. Comparative study of RNA-seq-and microarray-derived coexpression networks in Arabidopsis thaliana. Bioinformatics 2013; 29(6): 717-24.
18. de Matos Simoes R, Dalleau S, Williamson KE and Emmert-Streib F. Urothelial cancer gene regulatory networks inferred from large-scale RNAseq, Bead and Oligo gene expression data. BMC Syst Biol 2015; 9: 21.
19. Altay G, Altay N, Neal D. Global assessment of network inference algorithms based on available literature of gene/protein interactions. Turk J Biol 2013; 37(5): 547-55.
20. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Favera RD, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics 2006; 7(1): S7.
21. Altay G, Emmert-Streib F. Inferring the conservative causal core of gene regulatory networks. BMC Systems Biology 2010; 4(1): 132.
22. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, et al. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol 2007; 5(1): e8.
23. Meyer PE, Kontos K, Lafitte F, Bontempi G. Information-theoretic inference of large transcriptional regulatory networks. EURASIP J Bioinform Syst Biol 2007; 2007(1): 79879.
24. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008; 9(1): 559.
25. Meyer PE, Lafitte F, Bontempi G. minet: AR/Bioconductor package for inferring large transcriptional networks using mutual information. BMC Bioinformatics 2008; 9(1): 461.
26. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003; 13(11): 2498-504.
27. Labhart P, Karmakar S, Salicru EM, Egan BS, Alexiadis V, O'Malley BW, et al. Identification of target genes in breast cancer cells directly regulated by the SRC-3/AIB1 coactivator. Proc Natl Acad Sci USA 2005; 102(5): 1339-44.
28. Nestal de Moraes G, Delbue D, Silva KL, Robaina MC, Khongkow P, Gomes AR, et al. FOXM1 targets XIAP and Survivin to modulate breast cancer survival and chemoresistance. Cell Signal 2015; 27(12): 2496-505.
29. Dawson SJ, Makretsov N, Blows FM, Driver KE, Provenzano E, Le Quesne J, et al. BCL2 in breast cancer: a favourable prognostic marker across molecular subtypes and independent of adjuvant therapy received. Br J Cancer 2010; 103(5): 668-75.
30. Dang CV. MYC on the path to cancer. Cell 2012; 149(1): 22-35.
31. Gerger A, Hofmann G, Langsenlehner U, Renner W, Weitzer W, Wehrschütz M, et al. Integrin alpha-2 and beta-3 gene polymorphisms and colorectal cancer risk. Int J Colorectal Dis 2009; 24(2): 159-63.

32. Wang Y, Kuramitsu Y, Ueno T, Suzuki N, Yoshino S, Iizuka N, et al. Glyoxalase I (GLO1) is up-regulated in pancreatic cancerous tissues compared with related non-cancerous tissues. Anticancer Res 2012; 32(8): 3219-22.
33. Hermann-Kleiter N, Klepsch V, Wallner S, Siegmund K, Klepsch S, Tuzlak S, et al. The nuclear orphan receptor NR2F6 is a central checkpoint for cancer immune surveillance. Cell rep 2015; 12(12): 2072-85.
34. Wu CH, Sahoo D, Arvanitis C, Bradon N, Dill DL, Felsher DW. Combined analysis of murine and human microarrays and ChIP analysis reveals genes associated with the ability of MYC to maintain tumorigenesis. PLoS Genet 2008; 4(6): e1000090.
35. Grützmann R, Pilarsky C, Ammerpohl O, Lüttges J, Böhme A, Sipos B, et al. Gene expression profiling of microdissected pancreatic ductal carcinomas using high-density DNA microarrays. Neoplasia 2004; 6(5): 611-22.
36. Kobberup S, Nyeng P, Juhl K, Hutton J, Jensen J. ETS-family genes in pancreatic development. Dev Dyn 2007; 236(11): 3100-10.
37. Wu X, Deng F, Li Y, Daniels G, Du X, Ren Q, et al. ACSL4 promotes prostate cancer growth, invasion and hormonal resistance. Oncotarget 2015; 6(42): 44849-63.
38. Rüenauver K, Menon R, Svensson MA, Carlsson J, Vogel W, Andrén O, et al. Prognostic significance of YWHAZ expression in localized prostate cancer. Prostate Cancer Prostatic Dis 2014; 17(4): 310-4.
39. He H, Dai F, Yu L, She X, Zhao Y, Jiang J, et al. Identification and characterization of nine novel human small GTPases showing variable expressions in liver cancer tissues. Gene Expression 2002; 10(5-6): 231-42.