Integrative Biological Network Analysis to Identify Shared Genes in Metabolic Disorders

Samet Tenekeci and Zerrin Isik

Abstract—Identification of common molecular mechanisms in interrelated diseases is essential for better prognoses and targeted therapies. However, complexity of metabolic pathways makes it difficult to discover common disease genes underlying metabolic disorders; and it requires more sophisticated bioinformatics models that combine different types of biological data and computational methods. Accordingly, we built an integrative network analysis model to identify shared disease genes in metabolic syndrome (MS), type 2 diabetes (T2D), and coronary artery disease (CAD). We constructed weighted gene co-expression networks by combining gene expression, protein-protein interaction, and gene ontology data from multiple sources. For 90 different configurations of disease networks, we detected the significant modules by using MCL, SPICi, and Linkcomm graph clustering algorithms. We also performed a comparative evaluation on disease modules to determine the best method providing the highest biological validity. By overlapping the disease modules, we identified 22 shared genes for MS–CAD and T2D–CAD. Moreover, 19 out of these genes were directly or indirectly associated with relevant diseases in the previous medical studies. This study does not only demonstrate the performance of different biological data sources and computational methods in disease-gene discovery, but also offers potential insights into common genetic mechanisms of the metabolic disorders.

Index Terms—Gene expression, gene ontology, gene-disease association, protein-protein interaction, metabolic syndrome, type 2 diabetes, coronary artery disease

1 Introduction

Metabolic disorders consist of a large set of inherited or acquired genetic diseases characterized by enzyme deficiencies that disrupt the normal metabolic process by causing abnormal chemical reactions in the body. Most of the metabolic disorders derive from metabolic syndrome (MS) which is a combination of pathological conditions including insulin resistance, hypertension, hyperlipidemia, and abdominal obesity. Patients with MS have three-fold increased risk of developing cardiovascular diseases (CVDs) and five-fold increased risk of developing diabetes mellitus [1], [2]. The International Diabetes Federation (IDF) estimates that 25 percent of the global population has MS [3], while 8.8 percent has diabetes [4], and the annual health-care expenditures exceed USD 727 billions, as of 2017 [5]. On the other hand, coronary artery disease (CAD), which accounts for 80 percent of all CVD diagnoses, is referred as the leading cause of death globally [6].

Understanding the underlying molecular mechanisms of metabolic disorders is crucial not only to reveal the cause of the disease but also to design targeted drug therapies. In this respect, identifying the shared disease genes for multiple metabolic disorders provides a potential insight into the state of the disease while enhancing the accuracy of the early diagnoses. However, identifying the common genes or protein complexes is a challenging task because of the fact that the metabolic disorders act on so many pathways that produce a large number of potential risk factors [7]. Therefore, it requires us to develop some integrative bioinformatics models combining multiple biological data sources and computational methods to establish valid and reliable results using reasonable amount of resources.

Identifying the shared genes by utilization of an integrative model requires the same essential data-mining procedures that a conventional disease-gene discovery practice requires, such as preprocessing, mapping, integration, feature selection, clustering, classification, and validation. In addition, parameterization and selection of the computational methods should be handled carefully since the performance of the model is highly dependent on the data sets and computational methods used. Thus, it is usually needed to perform a comparative evaluation on as many methods and configurations as possible.

In this paper, we suggest an integrative and comparative bioinformatics approach that combines multiple biological data sources and computational methods to identify shared disease genes in MS, T2D, and CAD. First, we construct weighted gene co-expression networks (WGCNs) for each disease group by integrating peripheral blood gene expression data of 29 subjects, protein-protein interaction (PPI) networks, and Gene Ontology (GO) information. Then, we execute different network clustering algorithms to cluster 90 disease networks, which are constructed using different parameters, and detected the disease modules for each disease group. After comparatively evaluating the clustering results, we overlap the networks establishing the highest

- *S. Tenekeci is with the Department of Computer Engineering, Izmir Institute of Technology, 35430 Izmir, Turkey. E-mail: samettenekeci@iyte.edu.tr.*
- *Z. Isik is with the Department of Computer Engineering, Dokuz Eylul University, 35370 Izmir, Turkey. E-mail: zerrin@cs.deu.edu.tr.*

Manuscript received 29 Jan. 2020; revised 22 Apr. 2020; accepted 2 May 2020. Date of publication 8 May 2020; date of current version 3 Feb. 2022.
(Corresponding author: Samet Tenekeci.)
Digital Object Identifier no. 10.1109/TCBB.2020.2993301

1545-5963 © 2020 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: ULAKBIM UASL - IZMIR YUKSEK TEKNOLOJI ENSTITUSU. Downloaded on July 27,2022 at 13:01:15 UTC from IEEE Xplore. Restrictions apply.

biological homogeneity and stability, and thus we obtained the most significant common disease modules. Finally, we validate the proposed gene-disease associations (GDAs) on several publicly available microarray data sets and the existing literature data in DisGeNET platform [8]. In doing so, we aim to explore novel disease genes shared between MS, T2D, and CAD, as well as to compare and evaluate the performance of different PPI networks, GO semantic similarity measures, and graph clustering algorithms, in disease-gene identification.

2 Related Work

2.1 Medical Studies on MS, T2D, and CAD

Over the past three decades, many medical researches have been carried out to reveal the relationship between metabolic disorders, and their results clearly demonstrated the tie between the MS, T2D, and CVDs. In 2001, Isomaa *et al.* [1] found that MS was present in 80 percent of subjects with T2D and the presence of MS increased the risk of coronary heart disease (CHD) three-fold and increased the risk of cardiovascular mortality and morbidity by 1.8-fold. In 2004, Grundy [9] described obesity-induced MS as a multidimensional risk factor for atherosclerotic cardiovascular disease (ASCVD) and T2D. Again in 2004, Grundy *et al.* [10] reported that in patients with MS, the risk of developing ASCVD increases at least twice, and the risk of developing T2D increases five times, regardless of gender. In 2005, Wilson *et al.* [11] observed that MS accounts for up to one third of CVD in men and approximately half of new T2D over eight years of follow-up. Many other studies [12], [13], [14], [15], [16] demonstrated the association and parallel incidence of MS, T2D, and CAD.

2.2 Computational Studies on MS, T2D, and CAD

Despite the biological complexity of the problem, several bioinformatics approaches have emerged to reveal common molecular mechanisms of MS, T2D, and CAD, by means of the recent developments in the post-genomic era. As well as these approaches get use of biological interaction data, gene expression data, sequence data, or GO information, they also combined multiple data sources to improve the integrity and reliability of the results. With an integrative analysis on biological pathways and networks, Chan *et al.* [17] discovered multiple biological pathways and key regulatory genes involved in CVD and T2D development. Ko *et al.* [18] proposed a novel approach that utilizes underlying molecular pathways and common disease-related genes to identify comorbid diseases through molecular interaction networks. Liu *et al.* [19] performed a weighted gene co-expression network analysis (WGCNA) to identify specific hub genes and modules associated with CAD; and they associated 3711 genes in 21 modules with CAD. Shu *et al.* [20] conducted a broad integrative analysis based on five multi-ethnic genome-wide association studies; and they identified the common disease sub-networks and metabolic pathways in T2D and CVD. Zhao *et al.* [21] performed a genome-wide study on multiple ancestry groups including 265,678 T2D and 260,365 CHD subjects; and they reported new genetic loci that are shared by CHD and T2D.

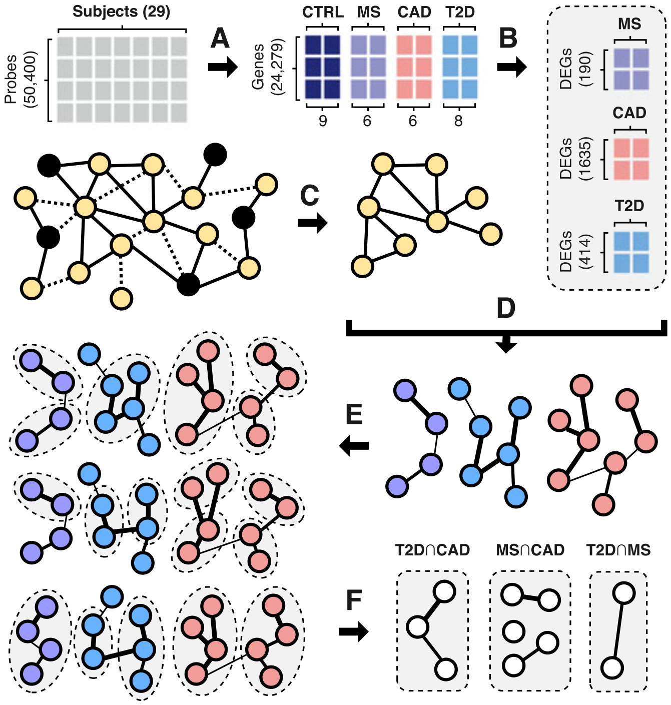

Fig. 1. Detection of common disease genes in six steps. (A) Preprocessing, normalization, and aggregation of gene expression data. (B) Detection of DEGs. (C) Preprocessing and filtering of PPI data. (D) Construction of disease networks by integrating gene expression, GO, and PPI data. (E) Identification of disease modules by use of different clustering algorithms and network configurations. (F) Determination of the best method and identification of common significant genes.

On the other hand, these studies mostly focused on exploring the genetic similarities or differences between a pair of diseases rather than examining many-to-many relationships in a disease set. Besides, they typically rely on single computational method and validation strategy even though they diversify the utilized data sources. In this study, we propose a new pipeline to integrate multiple biological data sources, computational methods, and validation metrics for extensive analysis of many-to-many relationships between multiple diseases. Besides, we comparatively evaluate the performance of these sources, methods, and metrics in disease gene discovery. Main novelty of the proposed system is integration of disease transcriptome data and gene semantic similarity into PPI networks to obtain more reliable gene modules related with MS, T2D, and CAD as an outcome of network clustering algorithms.

3 Method

3.1 System Overview

We provide the pipeline of the proposed system as a pseudocode (Supplementary S1 available online). We also present a general overview in Fig. 1. In step A, after removing invalid and null rows, we apply a logarithmic normalization. Then, we filter out the duplicated rows (i.e., probes) and obtain 24,279 unique genes out of 50,400 probes. In step B, we detect DEGs by considering both significance scores of *t*-tests and fold-change (FC) values.

In step C, we reduce the size of the STRING [22] and INet [23] PPI networks. We first map the proteins in both networks to the gene symbols in our expression data set. By

Authorized licensed use limited to: ULAKBIM UASL - IZMIR YUKSEK TEKNOLOJI ENSTITUSU. Downloaded on July 27,2022 at 13:01:15 UTC from IEEE Xplore. Restrictions apply.

TABLE 1
Size of the STRING and INet topologies (I: initial, R: Reduced)

	STRING		INet	
	Nodes	Edges	Nodes	Edges
I	19,576	5,676,528	19,290	7,077,509
R	13,969 (61%)	568,020 (10%)	12,264 (64%)	710,660 (10%)

using the medium-confidence threshold, we shrink both networks to the 10 percent of their initial sizes. In step D, we construct disease networks by integrating the gene expression, GO, and PPI data. We map the protein identifiers of the STRING and INet networks to gene symbols of DEGs for each disease group. On the resulting topology, we assign GO semantic similarity scores as edge weights to generate biologically meaningful clusters. In step E, we separately execute the MCL [24], SPICi [25], and Linkcomm [26] graph clustering algorithms on each disease network and obtain the significant disease modules. In step F, we measure the biological homogeneity (BHI) and cluster stability of each network. Thus, we determine the best network construction parameters and clustering algorithm to be used in the extraction of common disease modules. Then, we identify the overlapping disease modules that are expected to include shared disease genes for MS, T2D, and CAD.

At the end of the six steps, we validate our results on the literature and other data sets. By utilizing the DisGeNET, we identify the functional relationships of the revealed genes with each other and their associations with relevant disease classes, and thereby present the biological validity. On the other hand, we partially reproduce the results on 17 different validation sets to test the model consistency.

3.2 Gene Expression Analysis

We obtain gene expression profiles for 29 subjects (CTRL=9, MS=6, T2D=8, CAD=6) from GSE23561 peripheral blood gene expression data set [27] which is publicly available in Gene Expression Omnibus (GEO). The series matrix file provided under GSE23561 consists of 50,400 oligonucleotide probes and their expression values for each subject. We first apply log 2-based normalization on the F635 median values. Then, we remove the invalid and null rows in the data set and perform median-based aggregation for multiple probes corresponding to the same gene by mapping probe identifiers to gene symbols (Symbol v12) using GPL10775 platform. As a result, we obtain 24,279 log transformed gene expression values for each one of the 29 samples.

We separately identify the DEGs for each disease group by utilizing both t-tests and FC values. First, we compare the mean of each disease group with the control group using two-sample t-tests and filter out the genes where $p > 0.05$. The number of the resulting genes are 307, 435, and 2679 for MS, T2D, and CAD, respectively. Then, we calculate the FC for each gene i by $FC_{(i)} = |m_{CTRL(i)} - m_{DISEASE(i)}|$, where $m_{CTRL(i)}$ denotes the mean of log-normalized expression values in control (CTRL) group and $m_{DISEASE(i)}$ denotes the mean of log-normalized expression values in a disease group (i.e., MS, T2D, or CAD). After applying the fold-change cutoff, $FC \geq 1$, the number of DEGs for MS, T2D, and CAD appears as 190, 414, and 1635, respectively.

TABLE 2
Number of Nodes and Edges in Each Disease Network

	STRING		INet	
	Nodes	Edges	Nodes	Edges
MS	34	25	22	21
T2D	106	107	53	41
CAD	786	3,786	608	5,645

3.3 Protein-Protein Interaction Network Analysis

To construct the disease network topologies, we utilize STRING [22] and INet [23], those are the databases providing the highest coverage in human PPINs. STRING is a functional protein association network that consists of 19,576 unique proteins and 5,676,528 unique interactions between them. On the other hand, INet is an integrated network including information of four weighted human gene association networks (FunCoup, HumanNet, HIPPIE, and STRING) and it consists of 19,290 unique genes and 7,077,509 unique interactions.

We perform mapping over official gene symbols of expression data set and protein identifiers of PPINs by use of org.Hs.eg.db package [28] in R-Bioconductor. By eliminating the unmapped nodes and the interactions between them, we filter out 27 percent of nodes and 34 percent of edges in the STRING; and 28 percent of nodes and 33 percent of edges in the INet.

Then, we choose the significant interactions by using a medium-confidence cutoff which is 0.400 for STRING and 0.175 for INet. The initial and resulting number of nodes and edges for each network are shown in Table 1. To obtain the disease networks, we also eliminate the nodes that are not represented in the DEG sets. The size of the remaining networks are listed in Table 2.

3.4 GO Similarity Analysis

On each disease network, we calculate GO semantic similarity scores for connected gene pairs using GoSemSim R package [29]. GOSemSim generates GO semantic similarity scores by using different similarity measures, combination methods, and orthogonal ontologies. The similarity measure is either one of four information content (IC) based method (Jiang, Lin, Resnik, Rel) or a graph-based method (Wang) that are used in determination of the semantic similarity of two GO terms. The combination strategy is one of the max, avg, rcmax, or best-match average (BMA) and needed to calculate overall semantic similarity score on all pairs of two GO term sets. On the other hand, the reference orthogonal ontologies can be biological process (BP), cellular component (CC), or molecular function (MF).

To combine GO terms, we select BMA which has been suggested as the best combination method in previous studies [30], [31]. On the other hand, we repeat our analyses for each type of similarity measure (Jiang, Lin, Resnik, Rel, Wang) and each type of orthogonal ontology (BP, CC, MF) to perform a comparative evaluation.

3.5 Clustering

Clustering algorithms can be classified according to 5 main features: (1) strategy: the algorithmic approach they based on, (2) weight consideration: operability on weighted networks,

Authorized licensed use limited to: ULAKBIM UASL - IZMIR YUKSEK TEKNOLOJI ENSTITUSU. Downloaded on July 27,2022 at 13:01:15 UTC from IEEE Xplore. Restrictions apply.

TABLE 3
Main Features of MCL, SPICi, and Linkcomm Clustering Algorithms

Algorithm	Strategy	Weight	Overlap	Coverage	Scalability
MCL	Flow-based	Yes	No	Partial	Low
SPICi	Heuristic	Yes	No	Partial	High
Linkcomm	Hierachical	Yes	Yes	Full	Low

(3) overlap: whether generating intersecting or discrete clusters, (4) coverage: whether including each element in a cluster or not, (5) scalability: ability to deal with large data. We run three clustering algorithms with different characteristics (MCL, SPICi, Linkcomm). MCL [24] is a stochastic flow-based clustering algorithm providing full coverage, SPICi [25] is a heuristic local clustering algorithm with high scalability, and Linkcomm [26] is a link similarity-based hierarchical graph-cut algorithm that generates overlapping clusters. On the other hand, each algorithm can operate on undirected weighted graphs. The main features of each algorithm are presented in Table 3.

3.5.1 Implementation Details

We use *MCL* [32] and *Linkcomm* [33] R packages, and *SPICi* [25] python library to generate clusters. We execute *mcl*, *getLinkCommunities*, and *spici* functions for MCL, Linkcomm, and SPICi, respectively. For each method, we set the input as an undirected and weighted graph without self loops. In SPICi, we set the minimum cluster density (*d*) to 0.5, the minimum support threshold (*g*) to 0.5, the minimum cluster size (*s*) to 2, and graph mode to 0 (sparse graph). On the other hand, we use the default parameters in MCL and Linkcomm. All executions are performed on a computer with Intel Core i5-4200U processor, 8 GB of RAM, and Ubuntu 18.04 operating system.

3.5.2 Execution Time and Memory Consumption

We present the average execution time and maximum memory consumption of each algorithm in Table 4. As it is expected, SPICi is significantly faster and memory efficient (i.e., scalable), especially in large networks such as CAD.

3.6 Validation

3.6.1 BHI

The Biological Homogeneity Index (BHI) indicates to what extent the genes placed in the same statistical clusters belong to the same functional classes [34]. It is a useful measure to compare the performance of a number of competing clustering algorithms applied to the same data set.

For two annotated genes x, y that belong to the same statistical cluster D, $C(x)$ is a functional class containing gene x. Similarly, $C(y)$ contains gene y. The indicator function $I(C(x) = C(y)) = 1$ if $C(x)$ and $C(y)$ match. As genes x and y are in the same statistical cluster, it is expected that the two functional classes to match. The biological similarity of the statistical clusters are

$$BHI = \frac{1}{k}\sum_{j=1}^{k}\frac{1}{n_j(n_j-1)}\sum_{x\neq y\in D_j} I(C(x)=C(y)),$$

TABLE 4
Average Execution Time (t: sec) and Memory Consumption (m: Mb)

		MCL		Linkcomm		SPICi	
		t	m	t	m	t	m
STRING	MS	0.2	1.2	1.6	0.1	0.006	3.7
	T2D	1.2	19.2	2.1	2.2	0.005	3.8
	CAD	37.4	2298	32.6	4289	0.02	4.8
INet	MS	0.8	0.9	1.5	0.2	0.006	3.7
	T2D	1.2	4.6	1.5	0.3	0.006	3.7
	CAD	28.5	1642	331	609	0.02	5.2

where k is the number of statistical clusters, $n_j = n(D_j \cap C)$ is the number of annotated genes in cluster D_j, and $n(A)$ is size of any set A.

3.6.2 Stability of Disease Modules

In addition to biological homogeneity, we also evaluate the stability of the disease modules obtained. We adapted for our purposes the method offered by Hopcroft *et al.* to measure the stability of network communities [35]. We obtain 20 perturbed networks of each disease network by randomly removing 5 percent of all its nodes and the edges connected to these nodes. Here, the selected nodes and edges are arranged to be *random* and *disjoint* with each other. Then, we run the clustering algorithms for each perturbed network and calculate the stability of the modules based on the average best-match values. Here, the best-match value for the cluster C' in the perturbed network P is defined as

$$bestmatch(C', P) = \max_{C\in O}\left\{\min\left(\frac{|C'\cap C|}{|C|}, \frac{|C'\cap C|}{|C'|}\right)\right\},$$

where C is a cluster in the original network O. The average best-match value of P with respect to O is the mean of best-match values obtained for each cluster in P.

3.6.3 DisGeNET

DisGeNET [8] is an integrative database collecting gene-disease association (GDA) data from different data sources including animal models (M), GWAS catalogs (I), expert curated repositories (C), and scientific literature (L). DisGeNET v6.0 contains 628,685 GDAs, between 17,549 genes and 24,166 diseases. It is a useful platform for the validation of computationally predicted disease genes through the investigation of existing knowledge about them. DisGeNET gives a score (S) for each GDA based on the supporting data sources, which is calculated by $S = C + M + I + L$ where

$$C = \begin{cases} 0.6 & \text{if } N_c > 2 \\ 0.5 & \text{if } N_c = 2 \\ 0.4 & \text{if } N_c = 1 \\ 0 & \text{otherwise} \end{cases} \qquad L = \begin{cases} 0.1 & \text{if } N_p > 9 \\ \frac{N_p}{100} & \text{if } N_p \leq 9 \end{cases}$$

$$M = \begin{cases} 0.2 & \text{if } N_m > 0 \\ 0 & \text{otherwise} \end{cases} \qquad I = \begin{cases} 0.1 & \text{if } N_i > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Authorized licensed use limited to: ULAKBIM UASL - IZMIR YUKSEK TEKNOLOJI ENSTITUSU. Downloaded on July 27,2022 at 13:01:15 UTC from IEEE Xplore. Restrictions apply.

TABLE 5
BHI Comparison

		MS	T2D	CAD	Average
Topology	STRING	0.308	0.449	0.467	0.408
	INet	**0.407**	**0.458**	**0.481**	**0.449**
Similarity	Jiang	0.425	0.492	0.477	0.465
	Lin	0.415	0.491	0.482	0.463
	Rel	0.441	0.491	0.481	0.471
	Resnik	0.300	0.326	**0.484**	0.370
	Wang	**0.456**	**0.492**	0.481	**0.476**
Ontology	BP	0.442	0.486	0.478	0.468
	CC	0.444	0.490	0.481	0.472
	MF	**0.480**	**0.498**	**0.483**	**0.487**
Clustering	MCL	0.441	0.493	0.467	0.467
	SPICi	**0.500**	**0.500**	**0.492**	**0.497**
	Linkcomm	**0.500**	**0.500**	0.490	**0.497**

Here, N_c is the number of curated sources, N_m is the number of model organisms, N_i is the number of inferential sources, and N_p is the number of publications supporting the GDA.

3.6.4 External Validation

We downloaded 17 publicly available microarray data sets (1 for MS, 9 for T2D, and 7 for CAD) that include gene expression analyses in adipose, muscle, blood, heart, and liver tissues. We could not find exactly the same setup with the original microarray data set (GSE23561), which covers all three diseases and samples were composed as peripheral blood gene expression. Therefore, we performed individual analysis of 17 data sets. We run our DEG detection pipeline on each data set and then examine the commonality of DEGs detected in each data set and disease-genes revealed by the original data set GSE23561. The details of each data set are given in Table 10 (Supplementary S2 available online).

4 Results

4.1 Biological Homogeneity and Stability Evaluations

N is the total number of disease networks constructed. It is a combination of diseases (D), topologies (T), similarity measures (M), orthogonal ontologies (O), and clustering algorithms (A). To select the most significant network construction method, we comparatively evaluate the biological homogeneity (BHI) achieved by each configuration.

We represent the average BHI scores in Table 5. In 92 out of 135 disease networks (68 percent) for each topology the INet outperforms the STRING network. On the other hand, the STRING is better in 19 cases (14 percent), while they are equally successful in 24 cases (18 percent).

GO similarity measures provide very close BHI scores. On MS and T2D, Wang outperforms the others (Jiang, Lin, Rel, and Resnik). Although Resnik generates slightly more homogeneous clusters on CAD, it is significantly worse on MS and T2D. Therefore, the Wang measure providing the best homogeneity on average, it is the best alternative

TABLE 6
Stability of the Disease Modules

	MS	T2D	CAD	Average
MCL	0.91 ± 0.09	0.96 ± 0.03	0.85 ± 0.03	0.91 ± 0.05
SPICi	0.96 ± 0.14	0.97 ± 0.04	0.83 ± 0.04	0.92 ± 0.07
Linkcomm	0.98 ± 0.03	0.95 ± 0.05	0.94 ± 0.01	0.96 ± 0.03

$$D:3\begin{cases}\mathit{MS}\\ \mathit{CAD}\\ \mathit{T2D}\end{cases}\qquad T:2\begin{cases}\mathit{STRING}\\ \mathit{INet}\end{cases}$$

$$M:5\begin{cases}\mathit{Jiang}\\ \mathit{Lin}\\ \mathit{Rel}\\ \mathit{Resnik}\\ \mathit{Wang}\end{cases}\qquad A:3\begin{cases}\mathit{MCL}\\ \mathit{Linkcomm}\\ \mathit{SPICi}\end{cases}\qquad O:3\begin{cases}\mathit{BP}\\ \mathit{CC}\\ \mathit{MF}\end{cases}.$$

On the disease networks constructed by use of the Wang measure and the INet topology, we also compare the orthogonal ontologies (BP, CC, and MF). The results show that MF outperforms the BP and CC.

Lastly, we run the clustering algorithms on the disease networks constructed by use of the INet topology, the Wang measure, and the MF ontology. The performance of SPICi and Linkcomm are very close and better than MCL. Thus, we select both SPICi and Linkcomm to verify the results and not to miss out any significant clusters revealed by one of the algorithms.

The stability of disease modules found by different clustering algorithms are given in Table 6. The modules identified by MCL and SPICi are highly stable, the best stable algorithm is the Linkcomm with an average best-match stability value of 0.96. As a conclusion, both BHI and stability scores are concordant for three clustering algorithms.

4.2 Discovered Disease Modules

We report the genes of the same disease module in a list format within the curly brackets. Linkcomm generated two modules for MS: {ANAPC2, VPS28, PCGF6, HCFC1} and {RSP9, ARPC1B, POLR2L, ANAPC2}, four modules for T2D: {SP1, POU2F1}, {HBD, ALAS2}, {PLEKHG5, RAC3, ARHGAP10}, {DYNC1I1, SPTBN2, RILP}, and 137 modules with 532 unique genes for CAD. On the other hand, SPICi generated one module for MS: {GSPT2, RPS9, POLR2L, ANAPC2}, six modules for T2D: {SP1, POU2F1}, {HBD, ALAS2}, {PELI3, MAP3K14}, {RAC3, NRBP1}, {TXNRD2, SAMM50, BCS1L}, and 68 modules with 296 unique genes for CAD.

4.3 Common Disease Modules

We overlapped a total of 143 disease modules produced by the Linkcomm algorithm for MS, T2D, and CAD. As a result, we identified two modules, {ARPC1B, ANAPC2, RPS9, POLR2L} and {VPS28, PCGF6, HCFC1, ANAPC2}, and 12 individual genes, (TNFSF13, TNFRSF13B, NFKBIB, FRG1, S100A8, ENTPD2, SIX3, LHX2, GSPT2, ISYNA1, COX5A, GLUD2), shared by MS–CAD (Fig. 2A). Besides, we found a single shared gene, SP1, alongside a shared module with two genes, {ALAS2, HBD}, for T2D–CAD (Fig. 2B).

Authorized licensed use limited to: ULAKBIM UASL - IZMIR YUKSEK TEKNOLOJI ENSTITUSU. Downloaded on July 27,2022 at 13:01:15 UTC from IEEE Xplore. Restrictions apply.

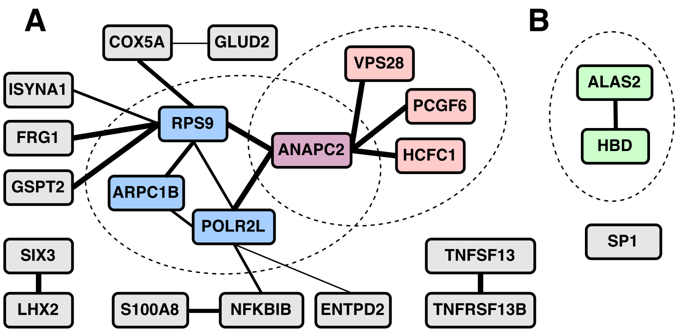

Fig. 2. Overlapping modules and genes of the Linkcomm clustering. (A) Two modules with seven genes and 12 individual genes are shared by MS–CAD pair. (B) One module with two genes (ALAS2, HBD) and the SP1 gene are shared by T2D–CAD pair. The gray nodes represent the common genes that are either unclustered or diversely clustered in each network. The edge thicknesses denote the GO similarity value of two genes. The modules with common nodes are shown in the same color.

Similarly, we overlapped a total of 75 disease modules produced by the SPICi algorithm for three diseases. As a result, we identified one module, {GSPT2, ANAPC2, RPS9, POLR2L}, and 15 individual genes, (TNFSF13, TNFRSF13B, NFKBIB, FRG1, S100A8, ENTPD2, SIX3, LHX2, ISYNA1, COX5A, GLUD2, VPS28, PCGF6, HCFC1, ANAPC2), shared by MS–CAD (Fig. 3A). As similar with Linkcomm, one common module, {ALAS2, HBD}, and one common gene, SP1, revealed for T2D–CAD pair (Fig. 3B).

Unlike MS–CAD and T2D–CAD, we could not detect a shared module for the MS–T2D using neither the SPICi nor the Linkcomm algorithms. Therefore, our analyses did not reveal a disease module shared by three diseases.

4.4 Evaluation of the Gene-Disease Associations

SPICi and Linkcomm identified 19 unique genes (ANAPC2, ARPC1B, COX5A, ENTPD2, FRG1, GLUD2, GSPT2, HCFC1, ISYNA1, LHX2, NFKBIB, PCGF6, POLR2L, RPS9, S100A8, SIX3, TNFRSF13B, TNFSF13, VPS28) shared between MS–CAD and three unique genes (ALAS2, HBD, SP1) shared between T2D–CAD. For these 22 genes, 1,136 GDAs have been reported in DisGeNET (Table 7).

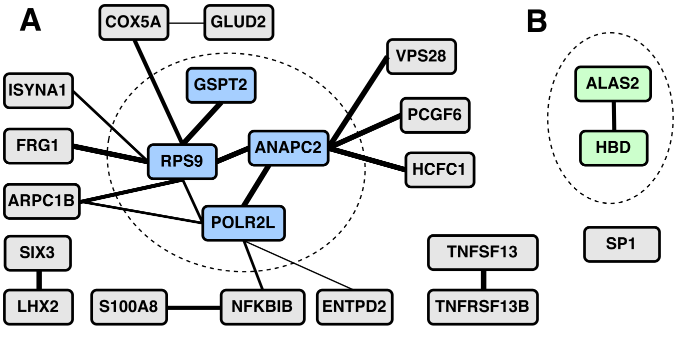

Fig. 3. Overlapping modules and genes of the SPICi clustering. (A) One module with four genes and 15 individual genes are shared by MS–CAD pair. (B) One module with two genes (ALAS2, HBD) and the SP1 gene are shared by T2D-CAD pair. The gray nodes represent the common genes that are either unclustered or diversely clustered in each network. The edge thicknesses denote the GO similarity value of two genes. The modules with common nodes are shown in the same color.

TABLE 7
Total Number of GDAs (N_{total}) Reported for Each Gene

Gene Symbol	Description	N_{total}
ANAPC2	APC2, WNT signaling pathway regulator	82
ARPC1B	actin related protein 2/3 complex subunit 1B	6
COX5A	cytochrome c oxidase subunit 5A	75
ENTPD2	ectonucleoside triphosphate diphosph 2	12
FRG1	FSHD region gene 1	47
GLUD2	glutamate dehydrogenase 2	5
GSPT2	G1 to S phase transition 2	4
HCFC1	host cell factor C1	43
ISYNA1	inositol-3-phosphate synthase 1	95
LHX2	LIM homeobox 2	16
NFKBIB	NFKB inhibitor beta	12
PCGF6	polycomb group ring finger 6	1
POLR2L	RNA Polymerase II Subunit L	0
RPS9	ribosomal protein S9	2
S100A8	S100 calcium binding protein A8	199
SIX3	SIX homeobox 3	84
TNFRSF13B	TNF receptor superfamily member 13B	118
TNFSF13	TNF superfamily member 13	90
VPS28	VPS28, ESCRT-I subunit	1
ALAS2	5′-aminolevulinate synthase 2	59
HBD	hemoglobin subunit delta	26
SP1	Sp1 transcription factor	159
TOTAL		**1136**

In nutritional and metabolic diseases (NMD) class, we identified 11 genes (APC2, COX5A, FRG1, HCFC1, ISYNA1, NFKBIB, S100A8, TNFSF13, ALAS2, HBD, SP1) with 57 GDAs reported by 82 articles (Table 8). In diabetes and diabetic complications (DIAB) sub class, we identified 4 genes (ISYNA1, S100A8, SIX3, SP1) with GDAs reported by 15 articles (Table 8). In cardiovascular diseases (CVD) class, we found 11 genes (ALAS2, APC2, COX5A, ENTPD2, ISYNA1, RPS9, S100A8, SIX3, SP1, TNFRSF13B, TNFSF13) with 60 GDAs reported by 88 articles (Table 8).

Five out of 19 genes (APC2, COX5A, ISYNA1, S100A8, TNFSF13) identified for MS–CAD were already associated with both NMD and CVD class diseases. On the other hand, three genes (FRG1, HCFC1, NFKBIB) were only associated with NMD, three genes (ENTPD2, RPS9, TNFRSF13B) were

TABLE 8
The Number of GDAs in Related Disease Classes (N_{nmd}, N_{diab}, N_{cvd}) and the Maximum DisGeNET Scores (S_{max})Reported for Each Gene

Gene Symbol	Disease Pair	N_{nmd}	N_{diab}	N_{cvd}	S_{max}
ANAPC2	MS–CAD	2	0	1	0.100
COX5A	MS–CAD	9	0	4	0.340
ENTPD2	MS–CAD	0	0	1	0.200
FRG1	MS–CAD	1	0	0	0.100
HCFC1	MS–CAD	4	0	0	0.100
ISYNA1	MS–CAD	6	6	13	0.030
RPS9	MS–CAD	0	0	1	0.010
NFKBIB	MS–CAD	1	0	0	0.010
S100A8	MS–CAD	14	4	13	0.040
SIX3	MS–CAD	0	1	2	0.100
TNFRSF13B	MS–CAD	0	0	6	0.450
TNFSF13	MS–CAD	2	0	9	0.060
ALAS2	T2D–CAD	13	0	3	0.600
HBD	T2D–CAD	1	0	0	0.100
SP1	T2D–CAD	4	1	7	0.320
TOTAL		57	12	60	

Authorized licensed use limited to: ULAKBIM UASL - IZMIR YUKSEK TEKNOLOJI ENSTITUSU. Downloaded on July 27,2022 at 13:01:15 UTC from IEEE Xplore. Restrictions apply.

TABLE 9
Matching of Our Results and the Previously Reported GDAs

Gene Symbol	Disease Pair	GDA reported class(es)	Matching
ANAPC2	MS–CAD	NMD + CVD	Full
ARPC1B	MS–CAD	–	None
COX5A	MS–CAD	NMD + CVD	Full
ENTPD2	MS–CAD	CVD	Half
FRG1	MS–CAD	NMD	Half
GLUD2	MS–CAD	–	None
GSPT2	MS–CAD	–	None
HCFC1	MS–CAD	NMD	Half
ISYNA1	MS–CAD	NMD + DIAB + CVD	Full
LHX2	MS–CAD	–	None
NFKBIB	MS–CAD	NMD	Half
PCGF6	MS–CAD	–	None
POLR2L	MS–CAD	–	None
RPS9	MS–CAD	CVD	Half
S100A8	MS–CAD	NMD + DIAB + CVD	Full
SIX3	MS–CAD	DIAB + CVD	Half
TNFRSF13B	MS–CAD	CVD	Half
TNFSF13	MS–CAD	NMD + CVD	Full
VPS28	MS–CAD	–	None
ALAS2	T2D–CAD	NMD + CVD	Half
HBD	T2D–CAD	NMD	None
SP1	T2D–CAD	NMD + DIAB + CVD	Full

only associated with CVD, and one gene (SIX3) was associated with DIAB and CVD. The remaining genes (ARPC1B, GLUD2, GSPT2, LHX2, PCGF6, POLR2L, VPS28) were suggested as novel for MS–CAD, since they were not previously associated with NMD or CVD (Table 9).

Among three genes (ALAS2, HBD, SP1) identified for T2D–CAD pair, only SP1 was previously associated with both DIAB and CVD. ALAS2 was associated with NMD and CVD, while HBD was associated with NMD (Table 9). Hence, we also suggest HBD as a novel gene for T2D–CAD.

Although our clustering algorithms could not reveal a disease module shared by all disease groups, three genes (ISYNA1, S100A8, SP1) that we identified were associated with NMD, DIAB, and CVD disease classes. Additionally, three genes (ALAS2, HBD, SIX3) are potentially shared by all disease groups, since they were associated with a complementary disease class in the DisGeNET (e.g., DIAB is a complementary class for a gene shared by MS–CAD). In pairwise associations, we obtained *full-matching* (i.e., previous association with both disease classes) for six genes and *half-matching* (i.e., previous association with one of the disease classes) for eight genes. In addition, we identified eight novel genes that have no previous association with any of the disease classes.

4.5 Functional and Relational Analysis on Novel Genes

To gain an insight into the functional and biological features of the eight novel disease genes (ARPC1B, GLUD2, GSPT2, HBD, LHX2, PCGF6, POLR2L, VPS28), we investigated their associations with non-metabolic disorders and their interactions with other genes that have been already associated with metabolic disorders.

ARPC1B encodes one out of seven subunits of the human Arp2/3 protein complex which has been implicated in the control of actin polymerization in cells. In addition, ARPC1B plays a major role in the regulation of the actin cytoskeleton and its deficiency causes platelet and immune system abnormalities [36]. On the other hand, ARPC1B is not previously associated with a metabolic disorder.

GLUD2 encodes an enzyme localized to the mitochondrion and acts as a homohexamer to recycle glutamate during neurotransmission. GLUD2 is associated with Parkinson disease [37], [38]. More importantly, its housekeeping isoform GLUD1 is clearly associated with several metabolic disorders and diabetic conditions [39], [40].

GSPT2 encodes a GTP-binding protein which has an essential role at the G1 to Sphase transition of the cell cycle. GSPT2 is associated with intellectual disability [41]. It is closely related to GSPT1 and shown to interact with PABPC1 [42]. However, none of these genes are previously associated with metabolic or cardiovascular diseases.

HBD and HBB genes are normally expressed in the adults and responsible from constitution of the hemoglobin. Mutations in HBD are associated with Deltathalassemia, an inherited blood disorder characterized by abnormal hemoglobin production [43]. On the other hand, HBB is associated with several CVDs [44], [45].

LHX2 encodes a protein that belongs to a large protein family, members of which carry the LIM domain and function as a transcriptional regulator. It is associated with neoplastic process, digestive system diseases, and rheumatoid arthritis [46]. LHX2 is also shown to interact with CITED2 [47] which is strongly associated with several heart diseases and defections [48], [49].

PCGF6 encodes a Polycomb group (PcG) protein, which acts as a master regulator to ensure embryonic stem cell development and differentiation [50]. PCGF6 is most closely related to PCGF2 that is known as a marker in breast cancer [51], [52]. However, neither of these genes are previously associated with a metabolic disorder.

POLR2L encodes a subunit of RNA polymerase II that is the polymerase responsible for synthesizing messenger RNA in eukaryotes, and it is shown to interact with POLR2A [53] which is associated with heart failure and cardiomyopathy [54].

VPS28 encodes a protein subunit of the ESCRT-I complex, which functions in the transportation and sorting of proteins into subcellular vesicles. Although there is not a GDA reported for VPS28 in the literature, VPS37C in the same subunit (ESCRT-I) is associated with rheumatoid arthritis and cardiometabolic disorders [55], [56].

Our DisGeNET and literature based analyses present that five out of eight genes (GLUD2, HBD, LHX2, POLR2L, VPS28) that we identified as novel disease genes for MS–CAD and T2D–CAD have some indirect associations with diseases in NMD and CVD class. On the other hand, there is no previous metabolic, diabetic, or cardiovascular disorder association reported for the remaining three genes (ARPC1B, GSPT2, PCGF6).

4.6 External Validation Results

We overlapped the 22 disease-genes found by the proposed system with the DEGs of 17 external validation data sets (1 for MS, 9 for T2D, and 7 for CAD) individually. We provide the validation data sets and detailed results in Table 10 (Supplementary S2 available online). When we consider overall results, we detected 2 matching genes (ALAS2, S100A8) for

Authorized licensed use limited to: ULAKBIM UASL - IZMIR YUKSEK TEKNOLOJI ENSTITUSU. Downloaded on July 27,2022 at 13:01:15 UTC from IEEE Xplore. Restrictions apply.

MS, 10 matches (ALAS2, ARPC1B, HBD, LHX2, NFKBIB, POLR2L, S100A8, SIX3, SP1, TNFSF13) for T2D, and 7 matches (ALAS2, FRG1, NFKBIB, S100A8, SIX3, SP1, TNFSF13) for CAD. Thus, we have externally validated 11 out of 22 GDAs (ALAS2, ARPC1B, FRG1, HBD, LHX2, NFKBIB, POLR2L, S100A8, SIX3, SP1, TNFSF13) which are discovery of the proposed system.

5 Conclusion

Identifying the shared disease genes for multiple diseases can be very useful for increasing the accuracy of prognosis and designing targeted therapies. In this respect, we propose a novel pipeline integrating multiple biological data sources, computational methods, and validation measures for analysis of many-to-many relationships between MS, T2D, and CAD.

We constructed 30 disease networks for each disease group by use of different PPIN topologies, orthogonal ontologies, and GO similarity measures. We clustered the revealed networks using three different clustering algorithms. Then, we evaluated the performance of each configuration in terms of stability and biological homogeneity achieved in the generated disease modules. As a result of the comparisons, we found that the highest BHI scores were obtained in networks constructed using the INet topology, the Wang similarity measure, and the MF ontology. On the other hand, the SPICi and Linkcomm algorithms were almost equally successful in generating biologically more homogeneous clusters, and better than MCL.

We identified 22 shared genes among MS–CAD and T2D–CAD pairs by overlapping the disease modules that are generated by use of the best configuration. Eleven of these genes (ALAS2, ARPC1B, FRG1, HBD, LHX2, NFKBIB, POLR2L, S100A8, SIX3, SP1, TNFSF13) were observed on different gene expression experiments related with targeted diseases, thus we validated novel disease genes by independent studies. Fourteen of them were previously associated with the nutritional and metabolic diseases (NMD), diabetes and diabetic complications (DIAB), and cardiovascular diseases (CVD) either with a full or partial matching. The remaining eight genes (ARPC1B, GLUD2, GSPT2, HBD, LHX2, PCGF6, POLR2L, VPS28) were determined as novel for the relevant diseases.

In order to interpret the functional and relational connections of the novel genes, we conducted two analyses: (1) the associations with non-metabolic disorders and (2) gene-gene interactions with external genes associated with NMD, DIAB, or CVD class diseases. As a result, we found that five of them (GLUD2, HBD, LHX2, POLR2L, VPS28) have some indirect associations with diseases in NMD and CVD class. However, no previous association with these diseases reported for three genes (ARPC1B, GSPT2, PCGF6).

Our study presented the performance of different biological data sources, computational methods, and validation metrics, in disease-gene discovery. Moreover, it provided some evidences that there are common disease genes underlying the MS, T2D, and CAD. Although, the proposed system is experimented on these three diseases, it can be implemented to analyze the relationships between different diseases with common genetic mechanisms.

Appendix

All data sets and source codes related with this study are available at https://github.com/smtnkc/go-cluster.

References

[1] B. Isomaa *et al.*, "Cardiovascular morbidity and mortality associated with the metabolic syndrome," *Diabetes Care*, vol. 24, no. 4, pp. 683–689, 2001.

[2] S. M. Grundy, "Metabolic syndrome pandemic," *Arteriosclerosis Thrombosis Vascular Biol.*, vol. 28, no. 4, pp. 629–636, 2008.

[3] S. O'neill and L. O'driscoll, "Metabolic syndrome: A closer look at the growing epidemic and its associated pathologies," *Obesity Rev.*, vol. 16, no. 1, pp. 1–12, 2015.

[4] K. Ogurtsova *et al.*, "IDF diabetes atlas: Global estimates for the prevalence of diabetes for 2015 and 2040," *Diabetes Res. Clin. Pract.*, vol. 128, pp. 40–50, 2017.

[5] N. Cho *et al.*, "IDF diabetes atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045, " *Diabetes Res. Clin. Pract.*, vol. 138, pp. 271–281, 2018.

[6] S. Mendis *et al.*, *Global Atlas on Cardiovascular Disease Prevention and Control*. Geneva, Switzerland: World Health Organization, 2011.

[7] R. La Forge, "Obesity, metabolic syndrome and cardiovascular disease," *IDEA Fitness J.*, vol. 1, no. 4, pp. 25–28, 2004.

[8] J. Pi ñero *et al.*, "DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants," *Nucleic Acids Res.*, vol. 45, 2017, Art. no. gkw943.

[9] S. M. Grundy, "Obesity, metabolic syndrome, and cardiovascular disease," *J. Clin. Endocrinol. Metab.*, vol. 89, no. 6, pp. 2595–2600, 2004.

[10] S. M. Grundy, B. Hansen, S. C. Smith Jr, J. I. Cleeman, R. A. Kahn, and C. Participants, "Clinical management of metabolic syndrome: Report of the american heart association/national heart, lung, and blood institute/american diabetes association conference on scientific issues related to management," *Circulation*, vol. 109, no. 4, pp. 551–556, 2004.

[11] P. W. Wilson, R. B. DAgostino, H. Parise, L. Sullivan, and J. B. Meigs, "Metabolic syndrome as a precursor of cardiovascular disease and type 2 diabetes mellitus," *Circulation*, vol. 112, no. 20, pp. 3066–3072, 2005.

[12] D. E. Laaksonen, H.-M. Lakka, L. K. Niskanen, G. A. Kaplan, J. T. Salonen, and T. A. Lakka, "Metabolic syndrome and development of diabetes mellitus: Application and validation of recently suggested definitions of the metabolic syndrome in a prospective cohort study," *Amer. J. Epidemiol.*, vol. 156, no. 11, pp. 1070–1077, 2002.

[13] G. Hu, Q. Qiao, J. Tuomilehto, B. Balkau, K. Borch-Johnsen , and K. Pyorala, "Prevalence of the metabolic syndrome and its relation to all-cause and cardiovascular mortality in nondiabetic european men and women," *Archives Internal Med.*, vol. 164, no. 10, pp. 1066–1076, 2004.

[14] M. C. Carr and J. D. Brunzell, "Abdominal obesity and dyslipidemia in the metabolic syndrome: Importance of type 2 diabetes and familial combined hyperlipidemia in coronary artery disease risk," *J. Clin. Endocrinol. Metab.*, vol. 89, no. 6, pp. 2601–2607, 2004.

[15] A. Galassi, K. Reynolds, and J. He, "Metabolic syndrome and risk of cardiovascular disease: A meta-analysis," *Amer. J. Med.*, vol. 119, no. 10, pp. 812–819, 2006.

[16] S. De Rosa , B. Arcidiacono, E. Chiefari, A. Brunetti, C. Indolfi, and D. P. Foti, "Type 2 diabetes mellitus and cardiovascular disease: Genetic and epigenetic links," *Front. Endocrinol.*, vol. 9, 2018, Art. no. 2.

[17] K. H. K. Chan *et al.*, "Shared molecular pathways and gene networks for cardiovascular disease and type 2 diabetes mellitus in women across diverse ethnicities," *Circ.: Cardiovas. Genetics*, vol. 7, no. 6, pp. 911–919, 2014.

[18] Y. Ko, M. Cho, J.-S. Lee, and J. Kim, "Identification of disease comorbidity through hidden molecular mechanisms," *Sci. Rep.*, vol. 6, 2016, Art. no. 39433.

[19] J. Liu, L. Jing, and X. Tu, "Weighted gene co-expression network analysis identifies specific modules and hub genes related to coronary artery disease," *BMC Cardiovas. Disorders*, vol. 16, no. 1, 2016, Art. no. 54.

[20] L. Shu *et al.*, "Shared genetic regulatory networks for cardiovascular disease and type 2 diabetes in multiple populations of diverse ethnicities in the united states," *PLoS Genetics*, vol. 13, no. 9, 2017, Art. no. e1007040.

Authorized licensed use limited to: ULAKBIM UASL - IZMIR YUKSEK TEKNOLOJI ENSTITUSU. Downloaded on July 27,2022 at 13:01:15 UTC from IEEE Xplore. Restrictions apply.

[21] W. Zhao *et al.*, "Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease," *Nat. Genetics*, vol. 49, no. 10, 2017, Art. no. 1450.

[22] D. Szklarczyk *et al.*, "STRING v11: Protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D607–D613, 2018.

[23] F. Yang, D. Wu, L. Lin, J. Yang, T. Yang, and J. Zhao, "The integration of weighted gene association networks based on information entropy," *PloS One*, vol. 12, no. 12, 2017, Art. no. e0190029.

[24] S. Dongen, "Graph clustering by flow simulation," PhD dissertation, Utrecht: University of Utrecht, 2000.

[25] P. Jiang and M. Singh, "SPICI: A fast clustering algorithm for large biological networks," *Bioinformatics*, vol. 26, no. 8, pp. 1105–1111, 2010.

[26] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, "Link communities reveal multiscale complexity in networks," *Nature*, vol. 466, no. 7307, 2010, Art. no. 761.

[27] B. L. Grayson, L. Wang, and T. M. Aune, "Peripheral blood gene expression profiles in metabolic syndrome, coronary artery disease and type 2 diabetes," *Genes Immunity*, vol. 12, no. 5, 2011, Art. no. 341.

[28] M. Carlson, "The org.hs.eg.db package," 2019. [Online]. Available: https://doi.org/doi:10.18129/B9.bioc.org.Hs.eg.db

[29] G. Yu, F. Li, Y. Qin, X. Bo, Y. Wu, and S. Wang, "GOSemSim: An R package for measuring semantic similarity among GO terms and gene products," *Bioinformatics*, vol. 26, no. 7, pp. 976–978, 2010.

[30] X. Wu, E. Pang, K. Lin, and Z.-M. Pei, "Improving the measurement of semantic similarity between gene ontology terms and gene products: Insights from an edge-and IC-based hybrid method," *PLoS One*, vol. 8, no. 5, 2013, Art. no. e66745.

[31] C. Pesquita, D. Faria, H. Bastos, A. E. Ferreira, A. O. Falcão, and F. M. Couto, "Metrics for GO based protein semantic similarity: A systematic evaluation," *BMC Bioinformatics*, vol. 9, no. 5. BioMed Central, 2008, Art. no. S4.

[32] M. L. Jäger, "The MCL package," 2015. [Online]. Available: https://cran.r-project.org/package=MCL

[33] A. T. Kalinka and P. Tomancak, "linkcomm: An R package for the generation, visualization, and analysis of link communities in networks of arbitrary size and type," *Bioinformatics*, vol. 27, no. 14, pp. 2011–2012, 2011.

[34] S. Datta and S. Datta, "Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes," *BMC Bioinformatics*, vol. 7, no. 1, 2006, Art. no. 397.

[35] J. Hopcroft, O. Khan, B. Kulis, and B. Selman, "Natural communities in large linked networks," in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2003, pp. 541–546.

[36] W. H. Kahr *et al.*, "Loss of the Arp2/3 complex component ARPC1B causes platelet abnormalities and predisposes to inflammatory disease," *Nat. Commun.*, vol. 8, 2017, Art. no. 14816.

[37] A. Plaitakis *et al.*, "Gain-of-function variant in GLUD2 glutamate dehydrogenase modifies Parkinson's disease onset," *Eur. J. Hum. Genetics*, vol. 18, no. 3, 2010, Art. no. 336.

[38] A. Plaitakis, I. Zaganas, and C. Spanaki, "Deregulation of glutamate dehydrogenase in human neurologic disorders," *J. Neurosci. Res.*, vol. 91, no. 8, pp. 1007–1017, 2013.

[39] C. A. Stanley, "Two genetic forms of hyperinsulinemic hypoglycemia caused by dysregulation of glutamate dehydrogenase," *Neurochem. Int.*, vol. 59, no. 4, pp. 465–472, 2011.

[40] C. Tran, V. Konstantopoulou, M. Mecjia, K. Perlman, S. Mercimek-Mahmutoglu, and J. B. Kronick, "Hyperinsulinemic hypoglycemia: Think of hyperinsulinism/hyperammonemia (HI/HA) syndrome caused by mutations in the GLUD1 gene," *J. Pediatric Endocrinol. Metab.*, vol. 28, no. 7/8, pp. 873–876, 2015.

[41] C. Grau *et al.*, "Xp11. 22 deletions encompassing CENPVL1, CENPVL2, MAGED1 and GSPT2 as a cause of syndromic X-linked intellectual disability," *PLoS One*, vol. 12, no. 4, 2017, Art. no. e0175962.

[42] S.-I. Hoshino, M. Imai, T. Kobayashi, N. Uchida, and T. Katada, "The eukaryotic polypeptide chain releasing factor (eRF3/GSPT) carrying the translation termination signal to the 3′-poly (A) tail of mRNA direct association of eRF3/GSPT with polyadenylate-binding protein," *J. Biol. Chem.*, vol. 274, no. 24, pp. 16 677–16 680, 1999.

[43] J. Vives-Corrons, M. Pujades, A. Miguel-García, A. Miguel-Sosa , and S. Cambiazzo, "Rapid detection of spanish (delta beta) zero-thalassemia deletion by polymerase chain reaction," *Blood*, vol. 80, no. 6, pp. 1582–1585, 1992.

[44] V. Dinakaran, A. Rathinavel, M. Pushpanathan, R. Sivakumar, P. Gunasekaran, and J. Rajendhran, "Elevated levels of circulating dna in cardiovascular disease patients: Metagenomic profiling of microbiome in the circulation," *PLoS One*, vol. 9, no. 8, 2014, Art. no. e105221.

[45] J. Makani *et al.*, "Genetics of fetal hemoglobin in tanzanian and british patients with sickle cell anemia," *Blood*, vol. 117, no. 4, pp. 1390–1392, 2011.

[46] C. Galligan, E. Baig, V. Bykerk, E. Keystone, and E. Fish, "Distinctive gene expression signatures in rheumatoid arthritis synovial tissue fibroblast cells: Correlates with disease activity," *Genes Immunity*, vol. 8, no. 6, 2007, Art. no. 480.

[47] D. J. Glenn and R. A. Maurer, "MRG1 binds to the LIM domain of Lhx2 and may function as a coactivator to stimulate glycoprotein hormone α-subunit gene expression," *J. Biol. Chem.*, vol. 274, no. 51, pp. 36 159–36 167, 1999.

[48] S. Sperling *et al.*, "Identification and functional analysis of cited2 mutations in patients with congenital heart defects," *Hum. Mutation*, vol. 26, no. 6, pp. 575–582, 2005.

[49] D. Su *et al.*, "Down-regulation of EBAF in the heart with ventricular septal defects and its regulation by histone acetyltransferase p300 and transcription factors smad2 and cited2," *Biochimica et Biophysica Acta (BBA)-Mol. Basis Disease*, vol. 1832, no. 12, pp. 2145–2152, 2013.

[50] W. Zhao *et al.*, "Essential role for polycomb group protein Pcgf6 in embryonic stem cell maintenance and a noncanonical polycomb repressive complex 1 (PRC1) integrity," *J. Biol. Chem.*, vol. 292, no. 7, pp. 2773–2784, 2017.

[51] J.-Y. Lee *et al.*, "Loss of the polycomb protein Mel-18 enhances the epithelial–mesenchymal transition by ZEB1 and ZEB2 expression through the downregulation of miR-205 in breast cancer," *Oncogene*, vol. 33, no. 10, 2014, Art. no. 1325.

[52] M. L. Riis, T. Lüders, A.-J. Nesbakken, H. S. Vollan, V. Kristensen, and I. R. Bukholm, "Expression of BMI-1 and Mel-18 in breast tissue-a diagnostic marker in patients with breast cancer," *BMC Cancer*, vol. 10, no. 1, 2010, Art. no. 686.

[53] J. Acker, M. de Graaff, I. Cheynel, V. Khazak, C. Kedinger, and M. Vigneron, "Interactions between the human RNA polymerase II subunits," *J. Biol. Chem.*, vol. 272, no. 27, pp. 16 815–16 821, 1997.

[54] T. Brattelid, L. H. Winer, F. O. Levy, K. Liestøl, O. M. Sejersted, and K. B. Andersson, "Reference gene alternatives to gapdh in rodent and human heart failure gene expression studies," *BMC Mol. Biol.*, vol. 11, no. 1, 2010, Art. no. 22.

[55] S. Eyre *et al.*, "High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis," *Nat. Genetics*, vol. 44, no. 12, 2012, Art. no. 1336.

[56] J. Kettunen *et al.*, "Genome-wide association study identifies multiple loci influencing human serum metabolite levels," *Nat. Genetics*, vol. 44, no. 3, 2012, Art. no. 269.

Samet Tenekeci received the BS and MS degrees in computer engineering from Izmir University, Izmir, Turkey, and Dokuz Eylul University, İzmir, Turkey, respectively. He worked as a research assistant with the Department of Computer Engineering, Izmir University. In 2018, he joined the Izmir Institute of Technology, where he is a research assistant currently working toward the PhD degree. His research interests include computational biology, bioinformatics, data mining, and machine learning.

Zerrin Isik received the BS degree in computer engineering from Dokuz Eylul University, İzmir, Turkey, the MS degree in computer science and engineering from Sabanci University, Tuzla, Turkey, and the PhD degree in computer engineering from Middle East Technical University, Ankara, Turkey. She worked as a postdoctoral researcher with the Biotechnology Center of TU Dresden, Germany. In 2014, she joined Dokuz Eylul University as an assistant professor. Her research interests include bioinformatics, machine learning, and data mining.

▷ **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: ULAKBIM UASL - IZMIR YUKSEK TEKNOLOJI ENSTITUSU. Downloaded on July 27,2022 at 13:01:15 UTC from IEEE Xplore. Restrictions apply.