RESEARCH ARTICLE OPEN ACCESS

Manuscript received September 18, 2023; revised September 20, 2023; accepted November 12, 2023; date of publication Nov. 12, 2023
Digital Object Identifier (**DOI**): https://doi.org/10.35882/ijeeemi.v5i4.335

How to cite: Angga Aditya Permana, Muhammad Fahrury Romdendine, Analekta Tiara Perdana, “Graph Analysis for the Discovery of Key Proteins in Type 2 Diabetes Mellitus”, Indonesian Journal of Electronics, Electromedical Engineering, and Medical Informatics, vol. 5, no. 4, pp. 201–209, November. 2023.

Graph Analysis for the Discovery of Key Proteins in Type 2 Diabetes Mellitus

Angga Aditya Permana[1], Muhammad Fahrury Romdendine[1], Analekta Tiara Perdana[2]
[1] Department of Informatics, Faculty of Engineering and Informatics, Universitas Multimedia Nusantara, Indonesia
[2] Department of Biology, Faculty of Science, Universitas Islam Negeri Sultan Maulana Hasanuddin Banten, Indonesia

Corresponding author: Muhammad Fahrury Romdendine (e-mail: muhammad.romdendine@lecturer.umn.ac.id).

ABSTRACT One of the metabolic diseases with a rising prevalence in Indonesia is Type 2 Diabetes Mellitus (T2DM). A collective effort from various sectors is required to seek solutions for T2DM. The proteomic approach, which focuses on proteins and their interactions related to T2DM, can be used to understand this condition. This research aims to model protein interactions associated with T2DM using a network graph, enabling the identification of key proteins that have the potential to serve as therapeutic targets or T2DM biomarkers. The graph analysis method used in this study involved four centrality measures: degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality. The validation method used to confirm the identified proteins is gene set enrichment analysis. The results obtained from the graph analysis using four centrality measures highlighted that seven out of 27 T2DM-related proteins are key proteins; these are: ABCC8, HNF4A, INS, KCNJ11, NEUROD1, PDX1, and SLC30A8. This study concludes that graph analysis on the interaction graph of T2DM-related proteins successfully identified key proteins that could potentially serve as T2DM biomarkers. Further medical investigation is imperative because computational identification alone is not sufficient to confirm the validity of the findings in this study.

INDEX TERMS Enrichment analysis, graph analysis, protein-protein interactions, type 2 diabetes mellitus.

I. INTRODUCTION

As a country in transition from a developing state to a developed one, Indonesia faces a multitude of health challenges among its population [1]. One of the most pressing health issues is Type 2 Diabetes Mellitus (T2DM), with Indonesia ranking among the countries with the highest prevalence globally [2], [3]. Data reveals an alarming 8.2% increase in the prevalence of T2DM in Indonesia up to 2023, surpassing the average prevalence observed in middle and lower-income countries [4]. Conversely, only one-third of the total T2DM patients in Indonesia receive proper treatment [4]. Addressing this issue demands a concerted effort from diverse stakeholders, involving multiple facets. One pivotal step forward is to conduct an in-depth investigation into the mechanisms underlying T2DM within the human body.

Type 2 Diabetes Mellitus (T2DM) is a metabolic disorder characterized by multiple phenotypes, including hyperglycemia, insulin resistance, and various comorbidities [5]. As a chronic and progressive disease, understanding the biological mechanisms underlying the development of T2DM is crucial [6]. Investigating these mechanisms often involves the study of proteins known to play a role in T2DM pathogenesis.

Protein-protein interaction analysis is a branch of proteomics that focuses on elucidating interactions between different proteins [7]. The STRING database compiles known protein interactions from various sources, such as text mining, co-expression matrix queries, metabolic pathway data, and other genomic resources [8]. These interactions among proteins associated with a specific disease can be represented as a graph network, a widely used approach for discovery analysis. Additionally, beyond proteomics research, methodologies like network pharmacology utilize graph or network theory to model interactions between drugs and biological systems [9].

Previous studies have employed the modeling of biological systems through interaction graphs, particularly in the context of protein-protein interactions. For instance, study by [10] employed this approach to model protein-protein interactions in the context of the infectious disease COVID-19, with the goal of identifying potential drug candidates. Similarly, study by [11] applied this methodology to group interacting proteins in Parkinson's disease, shedding light on the disease's underlying biological mechanisms. Study by [12] delved into understanding the patterns exhibited by antibiotic-resistant genes, while study by [13] combined graph analysis with graph-based machine learning convolutional neural networks to detect protein complexes. These diverse literature examples highlight the effectiveness of utilizing graph networks to model various interactions, particularly those between proteins, in advancing our understanding of disease mechanisms.

Our study aims to extend this approach to model the interactions of various proteins associated with Type 2 Diabetes Mellitus (T2DM) through interaction graph networks. Notably, there is a paucity of proteomic investigations that explore T2DM from the perspective of graph theory and analysis. The resultant graph is subjected to rigorous analysis, with the primary objective of identifying central proteins that potentially contribute to the onset or progression of T2DM.

In addition to graph analysis, this study also conducted gene set enrichment analysis (GSEA) on the key proteins identified through the graph analysis. GSEA serves the purpose of delving deeper into the mechanisms affected, particularly where these key proteins exert a significant influence. Broadly, GSEA can be defined as a series of analyses aimed at discerning which terms related to various biological mechanisms (including pathways, biological processes, and molecular functions) are enriched among a given set of genes or proteins [14]. This approach is widely employed in proteomics studies as it enhances insights pertaining to the biological mechanisms involving the proteins of interest [15].

The primary objective of this study is to identify pivotal proteins in the context of Type 2 Diabetes Mellitus (T2DM) using graph analysis. These identified key proteins hold the potential to serve as candidate biomarkers or therapeutic targets. The outcomes of this research are expected to contribute significantly to the prevention and treatment of T2DM, both within Indonesia and globally. The identified key proteins can serve as a foundational reference for further medical research endeavors concerning T2DM.

This paper is structured into four main sections. Following the introduction presented above, the materials and methods section provides a comprehensive explanation of data acquisition, data sources, the construction of interaction graphs, graph analysis procedures, and the employed GSEA methodology. Subsequently, the results and discussions section presents the findings at each stage of the methodology, accompanied by an in-depth discussion and interpretation of these results. Finally, the paper concludes with a summarizing section that encapsulates the key insights drawn from this study.

The research conducted has made significant contributions to the understanding of Type 2 Diabetes Mellitus (T2DM). Firstly, it has identified key proteins associated with T2DM, suggesting their potential utility as biomarkers. This finding not only advances our knowledge of the disease but also offers a potential avenue for the development of diagnostic tools and targeted therapies. Additionally, the research underscores the importance of graph interaction modeling and analysis as a promising approach within the field of proteomics. By using network graphs to analyze protein interactions related to T2DM, this study has demonstrated the power of this method in unraveling the complex relationships within biological systems. This highlights the potential for graph-based approaches to play a pivotal role in future research and the development of innovative solutions in the study of metabolic diseases and beyond.

II. MATERIALS AND METODS

A. DATA ACQUISITIONS

The data required for this study encompass two main components: protein data pertaining to T2DM and information regarding interactions between proteins. Protein data linked to T2DM was sourced from The Human Protein Atlas database [16], employing the keyword "type 2 diabetes mellitus" for querying purposes. Subsequently, all proteins retrieved from the database were compiled for analysis.

Following the compilation of T2DM-related proteins, the next step involved the acquisition of protein interaction data from the STRING database [8]. The query executed in STRING consisted of the set of T2DM-related proteins obtained in the previous step. The interaction data provided by the STRING database was then retrieved for subsequent analysis.

B. DATA PREPARATIONS

The collected data then undergoes a preparation phase, which includes the initial steps for constructing the interaction graph. The first task involves data selection, wherein extraneous attributes are filtered out from the downloaded interaction data. This step is essential as protein-protein interaction graphs only necessitate the source node and target node information.

Following the removal of unnecessary attributes, a comma-separated value (CSV) file is generated, containing the relevant source node and target node information. In this context, all the T2DM-related proteins gathered during data acquisition serve as source nodes. Conversely, the target nodes consist of all proteins that interact with the source nodes within the set of T2DM-related proteins.

C. GRAPH CONSTRUCTION

The construction of the graph is executed using the interaction data that was previously prepared. The resulting graph takes the form of an undirected graph, as this accurately represents the bidirectional nature of protein interactions [8]. Each node within the graph corresponds to a protein related to T2DM, and the presence of an edge connecting two nodes signifies an interaction between the respective proteins.

D. GRAPH ANALYSIS

The constructed graph is subsequently subjected to analysis to identify key proteins. This analysis involves the computation of centrality measurements, which are utilized to quantify the importance of nodes within a graph or network [17]. The values of these centrality measures in graphs modeling interactions between proteins associated with T2DM are instrumental in ranking the significance of proteins within the biological mechanism of T2DM.

In this study, four centrality measures are employed: degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality. The equations used to calculate these four centrality measures was adopted from the work of [18]. Each centrality measure carries its own unique meaning and interpretation concerning the protein-protein interaction graph.

1) DEGREE CENTRALITY

Degree centrality serves as a metric for assessing the significance of a node within a graph, primarily determined by the quantity of edges linking to that node. Essentially, it quantifies the extent to which a node is intricately linked to other nodes within the network [19]. Degree centrality stands out as a frequently employed centrality metric in network analysis, proving especially valuable when pinpointing nodes with extensive connections in a network [20]. Equation (1) from [18] is used for measuring degree centrality of node *i* from a graph.

$$C_d(i) = \frac{\sum_{j=1}^{N} A_{ij}}{N-1} \tag{1}$$

where A_{ij} denotes the adjacent link between nodess *i* and node *j* and N is the total number of nodes.

Within protein-protein interaction graphs, degree centrality becomes a valuable tool for detecting proteins that exhibit extensive connections with other proteins in the network. These exceptionally connected proteins are commonly denoted as "hubs" and are believed to assume pivotal roles within the network. For instance, hubs may participate in the regulation of multiple pathways or function as connectors, linking disparate sections of the network togethers.

2) CLOSENESS CENTRALITY

Closeness centrality represents a metric that gauges the speed at which information can travel through a specific node to reach other nodes within a graph. It quantifies the degree of brevity in the shortest paths from a node to all other nodes present in the graoh. The closer a node is in terms of its proximity to all other nodes, the more elevated its closeness centrality rating. Closeness centrality holds particular relevance in signaling networks and frequently emerges as a crucial parameter when seeking potential key proteins [21]. Equation (2) from [18] is used for measuring closeness centrality of node *i*.

$$C_c(i) = \frac{\sum d_G(i,k)}{N-1} \tag{2}$$

where $d_G(i,k)$ denotes the shortest path (geodesic distance) from nodes *i* to nodes *k*.

3) BETWEENNESS CENTRALITY

Betweenness centrality serves as a metric for assessing centrality within a graph, primarily relying on the concept of shortest paths. It quantifies the degree to which a vertex acts as a bridge or intermediary along paths connecting other vertices. Vertices with high betweenness centrality may exert substantial influence within a graph due to their control over the flow of information between other nodes. Additionally, their removal from the graph can disrupt communications between other vertices to the greatest extent, as they lie on the largest number of paths through which messages travel. Equation (3) from [18] is used for measuring betweenness centrality of node *i*.

$$C_b(i) = \frac{\sum_{j,k \in V} \sigma_{jk}(i)}{\sigma_{jk}} \tag{3}$$

where $\sigma_{jk}(i)$ is the total number of pairs (*j*,*k*) with $j \neq k \neq i$ and between nodes *j* and *k* there exists a path passing through nodes *i* and σ_{jk} is the total number of paths from nodes *j* to nodes *k*.

4) EIGENVECTOR CENTRALITY

In the realm of graph theory, eigenvector centrality stands as a metric to gauge a node's influence within a graph. It allocates relative scores to all nodes in the graph, operating under the premise that connections to nodes with higher scores exert a greater impact on the score of the node in question than connections to nodes with lower scores. A heightened eigenvector score signifies that a node is linked to numerous other nodes that, in turn, possess elevated scores themselves. This concept reflects the idea that a node's centrality is influenced by both the quantity and quality of its connections within the graph. Equation (4) from [18] is used for measuring eigenvector centrality of node *i*.

$$e_i = \frac{1}{\lambda_{max}} \sum_{j=1}^{N} A_{ij} x_j, for\ i = 1,2,3,\ldots,N. \tag{4}$$

where $(x_1, x_2, \dots, x_N)^t$ denotes the eigenvector of the largest eigenvalue λ_{max} from the adjacency matrix. It is the weighted average of the scores x_i of all nodes connected to nodes i.

E. KEY PROTEINS IDENTIFICATION

The identification of key proteins within the interaction graph, comprising various proteins associated with T2DM, relies on the four previously calculated centrality measures. This determination involves ranking each node based on its corresponding centrality value. Beyond ranking, sub-graphs are also constructed, comprising proteins that attain centrality values exceeding the some thresholds.

F. GENE SET ENRICHMENT ANALYSIS

GSEA represents a pivotal analytical step in our quest to unravel the intricate web of biological intricacies within T2DM. In this process, we meticulously scrutinized a select cohort of proteins that emerged as conspicuously significant in prior steps. The profound essence of GSEA lies in its capacity to unveil the profound intricacies of the biological tapestry underpinning T2DM, delving deep into the biological processes, molecular mechanisms, and pathways that orchestrate the disease's progression [22]. Through GSEA, we transcend mere identification, striving to elucidate the very essence of these protein's roles in the larger biological narrative of T2DM, offering invaluable insights that may illuminate novel therapeutic avenues and foster a deeper understanding of this complex ailment.

GSEA was conducted using Enrichr [23]. Significant protein sets from the previous stage were input to Enrichr to obtain data related to biological processes, molecular mechanisms, and pathways from the queried protein sets. After the three types of data were obtained, visualization of the three types of data was carried out using CytoScape software [24].

III. RESULTS

A. DATA ACQUISITIONS AND PREPARATIONS

Data acquisition on THPA using the keyword "type 2 diabetes mellitus" resulted in a set of 27 T2DM-related proteins. These proteins are ABCC8, C4A, CAPN10, GFPT2, HNF4A, HSPA4, IDE, IL6, INS, KCNJ11, LEP, LRP5, MT-ND1, NEUROD1, OAS1, PDX1, PEA15, PPARGC1B, PTPRN, PTPRN2, RRAD, SH2B3, SHBG, SLC2A3, SLC30A8, WFS1, ZFP57.

The set of proteins is then queried to the STRING database to get the interaction data. The results obtained interaction data between 25 proteins out of 27 proteins. Two proteins that were not found were MT-ND1 and PEA15. The data obtained from STRING was then downloaded for the next stages. The obtained data are necessary for graph construction and analysis.

B. GRAPH CONSTRUCTION AND ANALYSIS

The interaction graph between proteins is transformed into an undirected graph, resulting in a graph comprising a total of 25 nodes and 61 edges. Each node corresponds to an individual protein related to T2DM, and each edge signifies an interaction between two proteins. The decision to represent the graph as undirected stems from the bidirectional nature of these protein interactions, ensuring a more accurate modeling approach. The presence of 61 edges indicates the existence of 61 interaction relationships among the 25 proteins associated with T2DM. FIGURE 1 provides a visual representation of the graph derived from the interactions between T2DM-related proteins. The constructed graph was analyzed and the results are presented in subsequent sub-section.

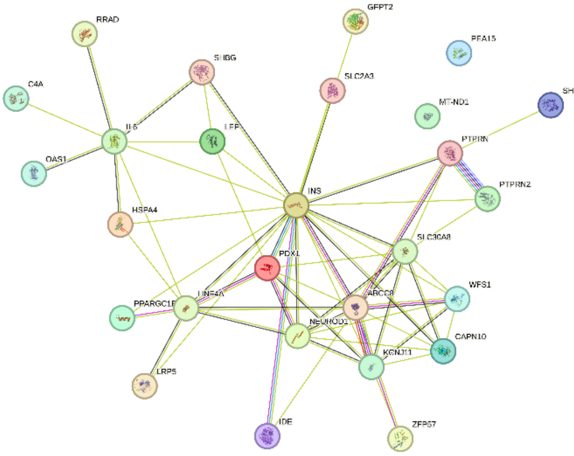

FIGURE 1. **Construction of protein-protein interaction graph resulted from visualization tools in STRING.**

C. KEY PROTEINS IDENTIFICATION

The identification of key proteins is based on the centrality values computed for each protein. The initial step involves ranking the proteins in descending order according to their centrality values. The calculated results for degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality for each protein are presented in TABLE 1.

To determine the key proteins, ranking is performed, starting with degree centrality. The top ten proteins with the highest degree centrality are identified as follows: INS, ABCC8, SLC30A8, NEUROD1, HNF4A, KCNJ11, IL6, PDX1, WFS1, and CAPN10. Creating a subgraph from these ten proteins with the highest degree centrality yields a graph comprising 10 nodes and 32 edges. The presence of 32 edges within the subgraph indicates that there are 32 interactions occurring among these ten proteins with the highest degree centrality. The visual representation of this subgraph is presented in FIGURE 2.

The protein with the highest degree centrality is INS or Insulin. If a protein has a high degree centrality, it means that many proteins interact with it, making it a vital protein in a protein interaction graph. Insulin is known to have an important role in T2DM disease. Cells in T2DM patients tend to be resistant to insulin even though insulin plays a role in regulating blood glucose levels [25]. This makes insulin the most vital protein when talking about T2DM. This study successfully confirmed the important role of insulin in T2DM through the perspective of graph analysis. The next ranking is based on closeness centrality. The ten proteins with the highest closeness centrality values are INS, ABCC8, HNF4A, SLC30A8, IL6, NEUROD1, KCNJ11, PDX1, WFS1, and CAPN10. The sub-set of proteins obtained is the same as when identifying using degree centrality, but there are differences in sequence compared to the sequence in degree centrality.

TABLE 1
Centrality measures of each protein

Protein/Gene	Degree Centrality	Closeness Centrality	Betweenness Centrality	Eigenvector Centrality
ABCC8	0.18033	0.02500	49.33333	0.35367
WFS1	0.09836	0.02174	0.00000	0.25633
IDE	0.03279	0.02000	0.00000	0.10294
NEUROD1	0.13115	0.02273	3.13333	0.31365
PTPRN	0.08197	0.02174	47.06667	0.16242
HNF4A	0.13115	0.02439	35.33333	0.23178
KCNJ11	0.13115	0.02273	16.26667	0.29674
PDX1	0.11475	0.02222	6.66667	0.26881
CAPN10	0.09836	0.02174	0.00000	0.25633
INS	0.31148	0.03448	324.93333	0.44241
SLC30A8	0.14754	0.02381	13.93333	0.31939
ZFP57	0.03279	0.01613	0.00000	0.08411
C4A	0.01639	0.01538	0.00000	0.01743
IL6	0.13115	0.02381	135.66667	0.13481
GFPT2	0.01639	0.01923	0.00000	0.05721
LRP5	0.03279	0.01961	0.00000	0.08718
HSPA4	0.04918	0.02128	0.00000	0.10462
PPARGC1B	0.03279	0.01961	0.00000	0.08718
LEP	0.06557	0.02174	3.66667	0.12108
SHBG	0.04918	0.02128	0.00000	0.09030
RRAD	0.01639	0.01538	0.00000	0.01743
OAS1	0.01639	0.01538	0.00000	0.01743
SLC2A3	0.01639	0.01923	0.00000	0.05721
PTPRN2	0.04918	0.02041	0.00000	0.11952
SH2B3	0.01639	0.01449	0.00000	0.02100

The subgraph formed from the ten proteins with the highest closeness centrality has a total of 10 nodes and a total of 27 edges. The same subgraph as the subgraph from the ranking results using degree centrality. Visualization of the subgraph formed from the ten proteins with the highest closeness centrality is presented in FIGURE 3.

Insulin emerges once again as the most crucial protein, as indicated by its high closeness centrality. This implies that insulin possesses the ability to efficiently transmit interaction information between proteins. Closeness centrality measures how rapidly a protein can establish connections with all other proteins in the graph, taking into account the shortest path length [26].

Subsequently, the identification of key proteins is conducted based on betweenness centrality values. In the protein-protein interaction graph, a protein can obtain a betweenness centrality value of zero, signifying the absence of any shortest paths passing through that particular protein. Among the T2DM-related proteins, those with non-zero betweenness centrality values include INS, IL6, ABCC8, PTPRN, HNF4A, KCNJ11, SLC30A8, PDX1, LEP, and NEUROD1.

Ten proteins exhibit non-null betweenness centrality values. The subgraph formed from these ten proteins in the protein-protein interaction network consists of ten nodes and 27 edges, as visualized in FIGURE 4. Identified key proteins based on betweenness centrality were subjected for GSEA and the results are presented in subsequent sub-section.

The identification of key proteins was further refined based on eigenvector centrality values. The top ten proteins with the highest eigenvector centrality values are INS, ABCC8, SLC30A8, NEUROD1, KCNJ11, PDX1, WFS1, CAPN10, HNF4A, and PTPRN. The interaction subgraph resulting from these ten proteins also comprises ten nodes, interconnected by 33 edges, as depicted in FIGURE 5.

Among these centrality measures, seven proteins—ABCC8, HNF4A, INS, KCNJ11, NEUROD1, PDX1, and SLC30A8—are consistently selected. Additionally, CAPN10, IL6, and WFS1 are chosen by all centrality measures except betweenness centrality, while PTPRN is exclusively identified by betweenness centrality and eigenvector centrality. The overlap of protein selection by each centrality measure is visually represented in a Venn diagram, presented as FIGURE 6.

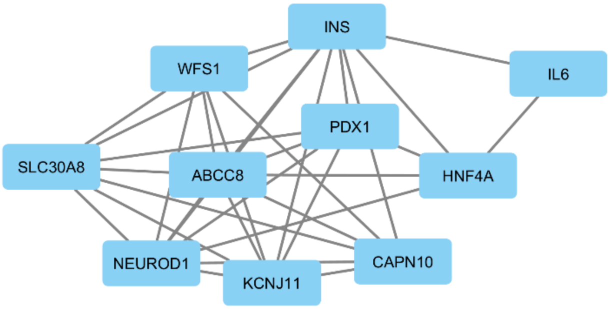

FIGURE 2. Sub-graph that models the interactions of ten proteins with highest degree centrality.

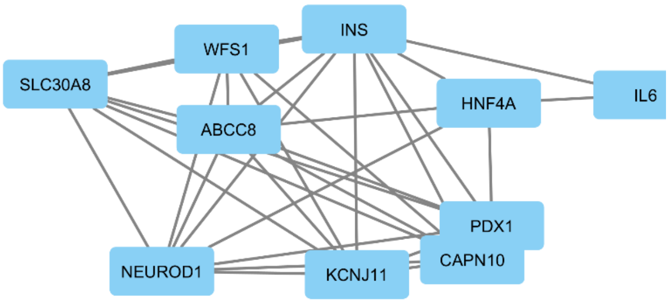

FIGURE 3. Sub-graph that models the interactions of ten proteins with highest closeness centrality.

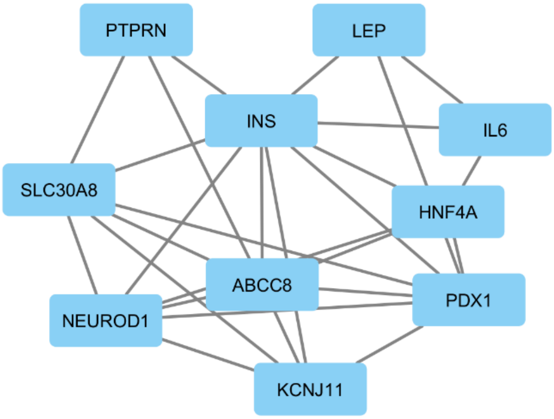

FIGURE 4. Sub-graph that models the interactions all proteins with non-zero betweenness centrality.

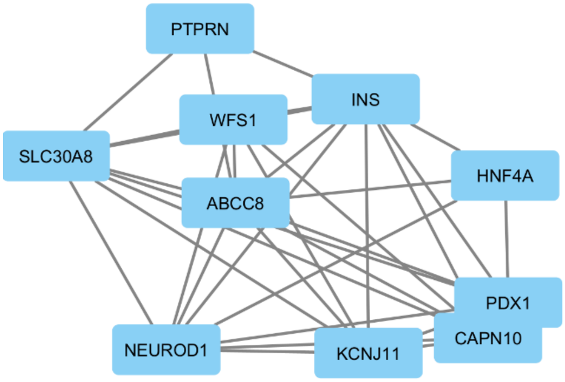

FIGURE 5. Sub-graph that models the interactions of ten proteins with highest eigenvector centrality.

D. GENE SET ENRICHMENT ANALYSIS

GSEA is conducted on all proteins that possess non-null betweenness centrality values, namely INS, IL6, ABCC8, PTPRN, HNF4A, KCNJ11, SLC30A8, PDX1, LEP, and NEUROD1. This approach is chosen due to the presence of proteins with betweenness centrality values equal to zero, making protein selection more straightforward. Moreover, proteins with non-zero betweenness centrality play a significant role in biological mechanisms, acting as hub proteins that facilitate interactions among two or more groups of proteins [27].

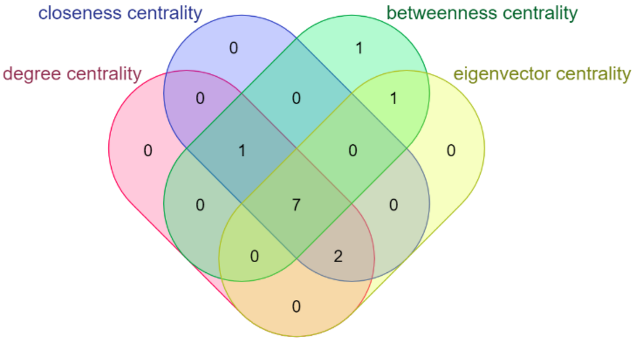

FIGURE 6. Venn diagram displaying proteins selected by all centrality measures.

1) KEGG PATHWAY

The first aspect under examination pertains to metabolic pathways. A metabolic pathway is a complex network of molecular interactions and reactions occurring within a cell or organism, serving a specific biological function [28]. Among the ten proteins with non-null betweenness centrality, the metabolic pathway most enriched is maturity onset diabetes of the young (MODY). This finding suggests that the key proteins identified in this study collectively contribute to the metabolic pathway associated with MODY. MODY represents a distinct type of diabetes characterized by gene mutations and typically manifests before the age of 25 [29]. While it differs from T2DM, MODY is frequently misdiagnosed as such. The study results indicate shared underlying mechanisms between the two conditions, warranting further clinical investigation. The bar chart in FIGURE 7 presents the top ten enriched metabolic pathways, sorted by p-value, corresponding to the ten proteins mentioned earlier.

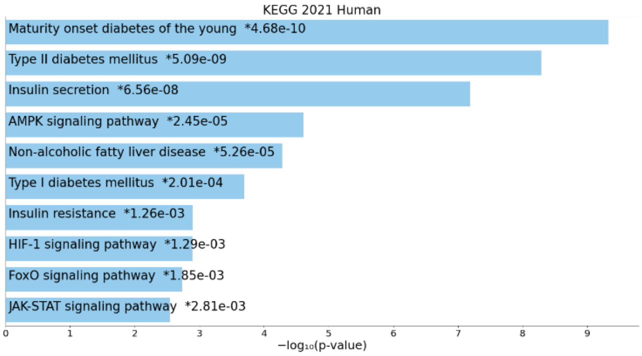

FIGURE 7. Bar chart of top enriched terms from the KEGG_2021_Human gene set library. The top 10 enriched terms for the input gene set are displayed based on the -log10(p-value), with the actual p-value shown next to each term. The term at the top has the most significant overlap with the input query gene set.

2) BIOLOGICAL PROCESS

Next, we investigate the biological processes enriched by the ten proteins mentioned earlier, with a focus on GO biological process terms. GO biological process terms describe a series of events within a cell or organism aimed at achieving specific biological objectives. For instance, 'cellular respiration' is a GO biological process term that outlines the process by which cells convert glucose into energy [30].

The most enriched GO biological process term is 'peptide hormone secretion'. This process involves the release of peptide hormones by endocrine cells into the human circulatory system [31]. Peptide hormone secretion is recognized for its influence on the development of T2DM, as certain peptide hormones are co-secreted with insulin, facilitating interactions between them [32]. The bar chart in FIGURE 8 illustrates the top ten enriched GO biological process terms associated with the ten selected proteins related to T2DM.

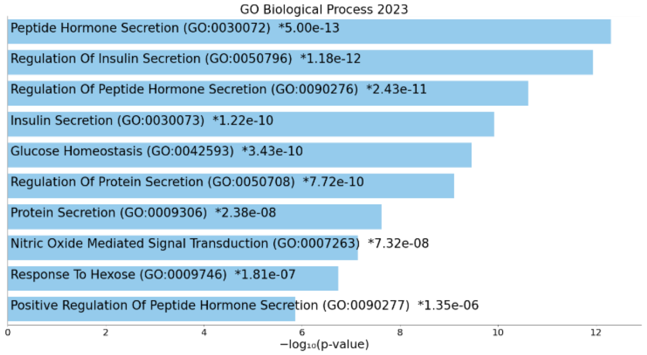

FIGURE 8. Bar chart of top enriched terms from the GO_Biological_Process_2023 gene set library. The top 10 enriched terms for the input gene set are displayed based on the -log10(p-value), with the actual p-value shown next to each term. The term at the top has the most significant overlap with the input query gene set.

3) MOLECULAR FUNCTION

The final aspect of GSEA examined in this study pertains to molecular function. The term 'molecular function' denotes a specific biochemical function of a protein, such as catalyzing a chemical reaction or binding to a particular molecule [33]. Understanding the molecular function of a protein is pivotal in comprehending its role within the biological mechanism of a disease, facilitating the identification of precise candidate therapeutic targets.

The most enriched GO molecular function term is 'receptor ligand activity'. This term signifies the function of a gene product that interacts with a receptor, initiating a cellular response [34]. 'Receptor ligand activity' encompasses a wide spectrum of ligand-receptor interactions, including those associated with cell signaling, growth factor activity, and cytokine activity [35]. The prominence of 'receptor ligand activity' among the enriched GO molecular function terms suggests that these proteins collectively play a role in influencing interactions between various other proteins to trigger cellular responses. This finding aligns with the fact that one of the phenotypes of T2DM is cellular resistance to insulin. FIGURE 9 illustrates the top ten enriched GO molecular function terms.

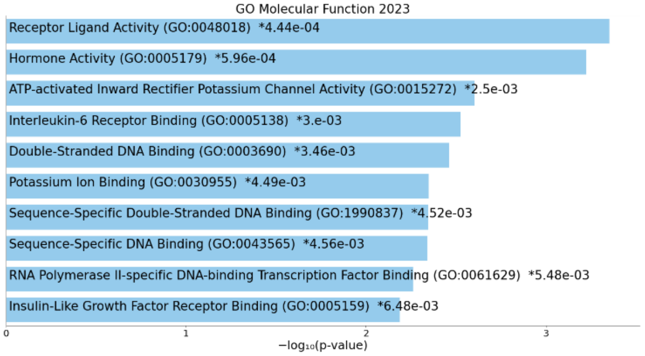

FIGURE 9. Bar chart of top enriched terms from the GO_Molecular_Function_2023 gene set library. The top 10 enriched terms for the input gene set are displayed based on the -log10(p-value), with the actual p-value shown next to each term. The term at the top has the most significant overlap with the input query gene set.

IV. DISCUSSIONS

T2DM as one of the deadly diseases cannot be separated from the involvement of various proteins that play a role in the biological system that causes the development of this disease. This study tries to reveal the role of various proteins involved in T2DM from the aspect of interactions between these proteins. We modeled the interactions between these proteins into the form of interaction graphs so that it can be seen whether graph analysis is able to identify key proteins in T2DM disease.

The identification is done using four types of centrality measures that are commonly used to determine the priority nodes in a graph or network. The four types of degree centrality have their own characteristics in determining which nodes are more prioritized than other nodes. Ranking based on the centrality measures value obtained by each node will produce a set of most prioritized proteins in the interaction graph that have different characteristics according to the type of centrality measures.

Degree centrality, a measure that quantifies the number of interactions between a protein and other proteins, highlights insulin as the most crucial protein. While this is not a novel revelation, it adds significant value to our modeling, as it can be regarded as a true positive. The presence of true positives is pivotal in the identification process, as it underscores the reliability and accuracy of our model [36].

Some key proteins identified through degree centrality share similarities with those identified through closeness centrality rankings. However, differences exist in the ranking order of proteins between these two centrality measures. For instance, HNF4A is ranked fifth in terms of having the most interactions (degree centrality) but climbs to the third position in terms of its effectiveness in communication and influence among proteins (closeness centrality). Closeness centrality, which quantifies how efficiently a protein communicates and influences other proteins in the interaction graph, also positions insulin as the top-ranked protein. This reaffirms the significance of insulin as a key true positive in this study.

Among the centrality measures applied in our graph analysis, betweenness centrality holds a unique significance, as it can assume a value of zero. In our study, it emerges as the most crucial metric for identifying key proteins related to T2DM. Notably, we observe an intriguing phenomenon: two proteins, PTPRN and LEP, which do not qualify as key proteins based on degree centrality or closeness centrality, are, nevertheless, identified as key proteins by betweenness centrality. This is a noteworthy finding because betweenness centrality quantifies a protein's pivotal role in bridging interactions between distinct protein clusters. Despite their limited number of interactions and comparatively less effective transmission of interaction information to other proteins, PTPRN and LEP play a critical role as bridges between two separate groups of interacting proteins. In practical terms, these proteins serve as key elements that, if targeted therapeutically, can influence the behavior of the two interconnected protein clusters. While they may not be explicitly designated as T2DM biomarkers, a study by [37] suggests that PTPRN has the potential to be a therapeutic target for T2DM patients with comorbid colorectal cancer. Another study [38] has linked polymorphisms in the LEP gene to an insulin resistance phenotype in a Malaysian population of T2DM patients, reinforcing the validity of their identification as true positives.

A noteworthy limitation of this study lies in the absence of subsequent in vitro or in vivo validation. The key proteins identified in this investigation cannot be unequivocally designated as biomarkers for T2DM or as effective therapeutic targets. This determination demands rigorous medical and clinical evaluation. However, the outcomes of this study hold the promise of serving as a critical foundation for the initiation of in vitro and in vivo validation processes. This anticipates that the validation procedure will be significantly expedited, as potential candidates for testing have already been identified through this research.

V. CONCLUSIONS

The aim of this study is to discover key proteins that have potentials to serve as T2DM biomarkers. The results showed that this study effectively modelled interactions among various T2DM-related proteins using interaction graphs. Through graph analysis, key proteins associated with T2DM were successfully identified. Seven proteins, namely ABCC8, HNF4A, INS, KCNJ11, NEUROD1, PDX1, and SLC30A8, emerged as the most promising key proteins, as they were consistently selected by all four centrality measures employed. These seven proteins hold significant potential as targets for T2DM drug development, with the potential to enhance therapeutic outcomes.

Furthermore, the analysis of ten proteins exhibiting non-null betweenness centrality values revealed their collective involvement in metabolic pathways, biological processes, and molecular functions, all of which displayed significant correlations with T2DM. Consequently, it can be inferred that this study has effectively identified key proteins associated with T2DM through graph analysis. However, computational identification alone, as conducted in this study, may not offer sufficient evidence to definitively classify the identified proteins as biomarkers for T2DM. Nevertheless, it serves as a valuable stepping stone for further research. Nevertheless, further investigation of these findings, especially from a clinical standpoint, is imperative to obtain medical validation.

ACKNOWLEDGMENT

The authors would like to thank Universitas Multimedia Nusantara for the support of this research work.

REFERENCES

[1] E. A. Finkelstein, J. Chay, and S. Bajpai, “The Economic Burden of Self-Reported and Undiagnosed Cardiovascular Diseases and Diabetes on Indonesian Households,” PLoS One, vol. 9, no. 6, p. e99572, Jun. 2014, doi: 10.1371/journal.pone.0099572.

[2] B. Arifin et al., “Health-related quality of life in Indonesian type 2 diabetes mellitus outpatients measured with the Bahasa version of EQ-5D,” Qual. Life Res., vol. 28, no. 5, pp. 1179–1190, May 2019, doi: 10.1007/s11136-019-02105-z.

[3] F. V. Ferdinand, J. Sebastian, and F. Natalia, “Predicting stroke, hypertension, and diabetes diseases based on individual characteristics,” ICIC Express Lett. Part B Appl., vol. 12, no. 8, pp. 723–731, 2021, doi: 10.24507/icicelb.12.08.723.

[4] H. Sofyan et al., “The state of diabetes care and obstacles to better care in Aceh, Indonesia: a mixed-methods study,” BMC Health Serv. Res., vol. 23, no. 1, p. 271, Mar. 2023, doi: 10.1186/s12913-023-09288-9.

[5] S. E. Kahn, M. E. Cooper, and S. Del Prato, “Pathophysiology and treatment of type 2 diabetes: perspectives on the past, present, and future.,” Lancet, vol. 383, no. 9922, pp. 1068–83, Mar. 2014, doi: 10.1016/S0140-6736(13)62154-6.

[6] R. A. DeFronzo et al., “Type 2 diabetes mellitus,” Nat. Rev. Dis. Prim., vol. 1, no. 15019, Jul. 2015, doi: 10.1038/nrdp.2015.19.

[7] Y. Du, J. Zhou, J. Fan, Z. Shen, and X. Chen, “Streamline proteomic approach for characterizing protein-protein interaction network in a RAD52 protein complex.,” J. Proteome Res., vol. 8, no. 5, pp. 2211–7, May 2009, doi: 10.1021/pr800662x.

[8] D. Szklarczyk et al., “The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest,” Nucleic Acids Res., vol. 51, no. D1, pp. D638–D646, Jan. 2023, doi: 10.1093/nar/gkac1000.

[9] S. Zhao and R. Iyengar, “Systems pharmacology: network analysis to identify multiscale mechanisms of drug action.,” Annu. Rev. Pharmacol. Toxicol., vol. 52, pp. 505–21, 2012, doi: 10.1146/annurev-pharmtox-010611-134520.

[10] M. Adhami, B. Sadeghi, A. Rezapour, A. A. Haghdoost, and H. MotieGhader, “Repurposing novel therapeutic candidate drugs for coronavirus disease-19 based on protein-protein interaction network analysis.,” BMC Biotechnol., vol. 21, no. 1, p. 22, Mar. 2021, doi: 10.1186/s12896-021-00680-z.

[11] R. Ferrari et al., “Stratification of candidate genes for Parkinson’s disease using weighted protein-protein interaction network

analysis.," BMC Genomics, vol. 19, no. 1, p. 452, Jun. 2018, doi: 10.1186/s12864-018-4804-9.

[12] N. A. Moumi, C. L. Brown, P. J. Vikesland, A. Pruden, and L. Zhang, "Protein-Protein Interaction Network Analysis Reveals Distinct Patterns of Antibiotic Resistance Genes," in 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Dec. 2022, pp. 73–76, doi: 10.1109/BIBM55620.2022.9995224.

[13] N. Zaki, H. Singh, and E. A. Mohamed, "Identifying Protein Complexes in Protein-Protein Interaction Data Using Graph Convolutional Network," IEEE Access, vol. 9, pp. 123717–123726, 2021, doi: 10.1109/ACCESS.2021.3110845.

[14] Y. Zhang, C.-Y. Li, W. Ge, and Y. Xiao, "Exploration of the Key Proteins in the Normal-Adenoma-Carcinoma Sequence of Colorectal Cancer Evolution Using In-Depth Quantitative Proteomics.," J. Oncol., vol. 2021, p. 5570058, 2021, doi: 10.1155/2021/5570058.

[15] J. H. Prieto, S. Koncarevic, S. K. Park, J. Yates, and K. Becker, "Large-scale differential proteome analysis in Plasmodium falciparum under drug treatment.," PLoS One, vol. 3, no. 12, p. e4098, 2008, doi: 10.1371/journal.pone.0004098.

[16] M. Uhlén et al., "A Human Protein Atlas for Normal and Cancer Tissues Based on Antibody Proteomics," Mol. Cell. Proteomics, vol. 4, no. 12, pp. 1920–1932, Dec. 2005, doi: 10.1074/mcp.M500279-MCP200.

[17] H. C. Rustamaji et al., "A network analysis to identify lung cancer comorbid diseases," Appl. Netw. Sci., vol. 7, no. 1, p. 30, Dec. 2022, doi: 10.1007/s41109-022-00466-y.

[18] S. L. Gan and M. A. Djauhari, "An Overall Centrality Measure : The Case of U . S Stock Market," Int. J. Basic Appl. Sci., vol. 12, no. 06, pp. 99–104, 2012.

[19] Z. A. Rachman, W. Maharani, and Adiwijaya, "The analysis and implementation of degree centrality in weighted graph in Social Network Analysis," in 2013 International Conference of Information and Communication Technology (ICoICT), Mar. 2013, pp. 72–76, doi: 10.1109/ICoICT.2013.6574552.

[20] O. Ledesma González, R. Merinero-Rodríguez, and J. I. Pulido-Fernández, "Tourist destination development and social network analysis: What does degree centrality contribute?," Int. J. Tour. Res., vol. 23, no. 4, pp. 652–666, Jul. 2021, doi: 10.1002/jtr.2432.

[21] M. Ashtiani et al., "A systematic survey of centrality measures for protein-protein interaction networks," BMC Syst. Biol., vol. 12, no. 1, p. 80, Dec. 2018, doi: 10.1186/s12918-018-0598-2.

[22] M. Neupane, J. N. Kiser, Bovine Respiratory Disease Complex Coordinated Agricultural Project Research Team, and H. L. Neibergs, "Gene set enrichment analysis of SNP data in dairy and beef cattle with bovine respiratory disease.," Anim. Genet., vol. 49, no. 6, pp. 527–538, Dec. 2018, doi: 10.1111/age.12718.

[23] M. V. Kuleshov et al., "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update," Nucleic Acids Res., vol. 44, no. W1, pp. W90–W97, Jul. 2016, doi: 10.1093/nar/gkw377.

[24] P. Shannon et al., "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks," Genome Res., vol. 13, no. 11, pp. 2498–2504, Nov. 2003, doi: 10.1101/gr.1239303.

[25] A. Gabryelska, F. F. Karuga, B. Szmyd, and P. Białasiewicz, "HIF-1α as a Mediator of Insulin Resistance, T2DM, and Its Complications: Potential Links With Obstructive Sleep Apnea.," Front. Physiol., vol. 11, p. 1035, 2020, doi: 10.3389/fphys.2020.01035.

[26] A. Sinha and H. A. Nagarajaram, "Nodes occupying central positions in human tissue specific PPI networks are enriched with many splice variants," Proteomics, vol. 14, no. 20, pp. 2242–2248, Oct. 2014, doi: 10.1002/pmic.201400249.

[27] A. Zito et al., "Gene Set Enrichment Analysis of Interaction Networks Weighted by Node Centrality.," Front. Genet., vol. 12, p. 577623, 2021, doi: 10.3389/fgene.2021.577623.

[28] G. R. Iyer et al., "Application of Differential Network Enrichment Analysis for Deciphering Metabolic Alterations.," Metabolites, vol. 10, no. 12, Nov. 2020, doi: 10.3390/metabo10120479.

[29] J. Taneera, P. Storm, and L. Groop, "Downregulation of type II diabetes mellitus and maturity onset diabetes of young pathways in human pancreatic islets from hyperglycemic donors.," J. Diabetes Res., vol. 2014, p. 237535, 2014, doi: 10.1155/2014/237535.

[30] K. Yang et al., "Exploring the Regulatory Mechanism of Hedysarum Multijugum Maxim.-Chuanxiong Rhizoma Compound on HIF-VEGF Pathway and Cerebral Ischemia-Reperfusion Injury's Biological Network Based on Systematic Pharmacology.," Front. Pharmacol., vol. 12, p. 601846, 2021, doi: 10.3389/fphar.2021.601846.

[31] D. J. Michael, H. Cai, W. Xiong, J. Ouyang, and R. H. Chow, "Mechanisms of peptide hormone secretion," Trends Endocrinol. Metab., vol. 17, no. 10, pp. 408–415, Dec. 2006, doi: 10.1016/j.tem.2006.10.011.

[32] S. V. Moelands, P. L. Lucassen, R. P. Akkermans, W. J. De Grauw, and F. A. Van de Laar, "Alpha-glucosidase inhibitors for prevention or delay of type 2 diabetes mellitus and its associated complications in people at increased risk of developing type 2 diabetes mellitus.," Cochrane database Syst. Rev., vol. 12, no. 12, p. CD005061, Dec. 2018, doi: 10.1002/14651858.CD005061.pub3.

[33] T. Sadlon et al., "Unravelling the molecular basis for regulatory T-cell plasticity and loss of function in disease.," Clin. Transl. Immunol., vol. 7, no. 2, p. e1011, 2018, doi: 10.1002/cti2.1011.

[34] The Gene Ontology Consortium, "The Gene Ontology Resource: 20 years and still GOing strong.," Nucleic Acids Res., vol. 47, no. D1, pp. D330–D338, Jan. 2019, doi: 10.1093/nar/gky1055.

[35] T. Nugroho and S. Prastowo, "Protein-to-protein interaction of genes responsible for the economic trait of Madura Cattle: an in silico analysis," IOP Conf. Ser. Earth Environ. Sci., vol. 1114, no. 1, p. 012084, Dec. 2022, doi: 10.1088/1755-1315/1114/1/012084.

[36] A. E. Ivanescu et al., "The importance of prediction model validation and assessment in obesity and nutrition research," Int. J. Obes., vol. 40, no. 6, pp. 887–894, Jun. 2016, doi: 10.1038/ijo.2015.214.

[37] M. Garranzo-Asensio et al., "Seroreactivity Against Tyrosine Phosphatase PTPRN Links Type 2 Diabetes and Colorectal Cancer and Identifies a Potential Diagnostic and Therapeutic Target," Diabetes, vol. 71, no. 3, pp. 497–510, Mar. 2022, doi: 10.2337/db20-1206.

[38] L. A. Ali, K. Jemon, N. Ab Latif, S. Abu Bakar, and S. S. Syed Alwi, "LEP G2548A Polymorphism is Associated with Increased Serum Leptin and Insulin Resistance among T2DM Malaysian Patients," BioMedicine, vol. 12, no. 3, pp. 31–39, Aug. 2022, doi: 10.37796/2211-8039.1326.